RELATIONSHIPS BETWEEN CHROMATIN FEATURES AND GENOME REGULATION

Przemyslaw Stempor

Girton College

The Gurdon Institute

Department of Genetics

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

December 2018

"Life would be tragic if it weren't funny."

Stephen Hawking

ABSTRACT

Relationships between chromatin features and genome regulation

Przemyslaw Stempor

Regulation of gene expression is an essential process for all living organisms. Transcriptional regulation, associated with chromatin, is governed by: (1) DNA sequence, which creates regulatory sites (promoters, enhancers and silencers), where sequence motifs and features (e. g. CpG) can attract transcription factors (TFs) and influence chromatin structure or RNA polymerase II (Pol II) binding, initiation and elongation; (2) non-sequence, epigenetic factors - histone modifications, TF binding, chromatin remodelling (histone placement, eviction and reconstitution), and non-coding RNA regulation. These factors interact with each other, creating a complex network of interactions. In this thesis I describe computational studies of heterochromatin factors in regulation of gene and repeat expression, an analysis of active regulatory elements, and global analyses of big datasets in *C. elegans*.

I first show that a team of heterochromatin factors - HPL-2/HP1, LIN-13, LIN-61, LET-418/Mi-2, and H3K9me2 histone methyltransferase MET-2/SETDB1 - collaborates with piRNA and nuclear RNAi pathways to silence repetitive elements and protect the germline. I also found that the TACBGTA motif is particularly enriched on repeats and heterochromatin factors binding sites, and that repeat elements are derepressed in the soma during normal *C. elegans* ageing.

I then describe the work on active regulatory regions. I show that CFP-1/CXXC1 binds CpG dense, nucleosome depleted promoters and, along SET-2, is required for H3K4me3 deposition at these loci. Using expression profiling I determined that the majority of CFP-1 binding targets are not significantly mis-regulated in *cfp-1* mutants, but are weakly upregulated in bulk analyses. I also show that CFP-1 functionally interacts with the Sin3S/HDAC complex. In *cfp-1* mutant I observed both loss and gain of SIN-3 binding, depending on chromatin context.

Finally, I performed a data driven study on a large collection of ChIP-seq profiles using non-parametric sparse factor analyses (NSFA) and compared it to other, unsupervised machine learning algorithms. This study uncovered interactions and structure in genomic datasets. In addition, I present a collection of computational tools and methods I developed to facilitate processing, storage, retrieval, annotation, and analyses of large datasets in genomics.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Professor Julie Ahringer, my thesis supervisor for interesting insights, countless valuable advices and continues support throughout the project. Her mentoring and profound biological knowledge proved to be essential for setting clear goals and navigating through embroiled filed of *C. elegans* biology. Further, I would like to thank professor Steve Russel, my project advisor for providing good feedback, especially at the planning phase. I would like to thank professor Zoubin Ghahramani, for directing me towards NSFA algorithm and hinting which textbooks I should read to start my adventure with machine learning.

I would like to thank all users of my software, in particular members of our laboratory and Gurdon Institute, for providing me with interesting design ideas, feedback about bugs, and everlasting motivation to keep developing better tools for scientific community.

I would also like thank all members of Ahringer Laboratory for support and creating excellent, challenging and friendly working environment. Next, I would like to acknowledge members of the Gurdon Institute for creating truly exceptional research environment, and really fun place to be. In my time here, I have met many extraordinary people, and made some great friendships.

Ultimately, I would like to thank my family and friend, in particular my parents, who have been supporting me through this incredible journey. Last, but not least, very special thanks go to Natalia Bulgakova for critically proofreading this thesis, always being a source of inspiration, amazing friend and true soulmate.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations and Acronyms

ChIP-seq – chromatin immunoprecipitation followed by sequencing

CV – coefficient of variation

DE – differential expression

DB – differential binding

DS – differential splicing

FDR – false discovery rate

GUI – graphical user interface

HC – heterochromatin

HCF – heterochromatin factor

HPC – high performance computing

HOT – highly occupied target

LFC – log fold change

NDR – nucleosome depleted region

O/E – observed over expected ratio

PCA – principal component analyses

SVD –singular value decomposition

TF – transcription factor

TSS – transcription start suite

TS – temperature sensitive

TTS – transcription termination site

UTR – untranslated region

WT – wild type

1 INTRODUCTION

1.1 Transcriptional regulation of gene expression

Regulation of gene expression is an essential process for all living organisms. The central dogma of molecular biology (Crick 1970) states that genes, encoded in DNA, are transcribed to mRNA and then translated to proteins. In multicellular organisms, all cells share, with rare exceptions, the same genetic material (DNA sequence), but the function and morphology differ strikingly between tissues. Much of this variability is controlled though gene expression, which can be regulated at the transcriptional and post-transcriptional levels. In my research, I am focusing on transcriptional regulation, which takes place in the nucleus and is strongly associated with chromatin.

Transcriptional regulation is governed by genomic sequence and auxiliary mechanisms:

(1) The DNA sequence is the primary factor creating regulatory sites such as promoters, enhancers and silencers. Certain sequence motifs and features (e. g. CpG density) can attract transcription factors (TFs), influence chromatin structure or directly influence RNA polymerase II (Pol II) binding and activity.

(2) Non-sequence factors include, among others, covalent histone tail modifications, transcription factor (TF) binding, histone variants, ATP-dependent chromatin

remodelling (histone placement, eviction and reconstitution), DNA methylation and non-coding RNA regulation (Teif & Rippe 2009).

1.2 The overview of *C. elegans* genome organisation

Caenorhabditis elegans has a compact genome (around 100 million base pairs), organised into 5 somatic chromosomes and a sex chromosome, with holocentric centromeres, plus mitochondrial DNA (Riddle *et al.* 1997). The autosomal chromosomes are structured into distal arm and central regions, with more actively transcribed genes being enriched in central regions, whereas distal arm regions are enriched for repeats and heterochromatin-like features (Liu *et al.* 2011b). General gene architecture is similar to that of other animals – genes consist of promoters, where the RNA Polymerase II complex binds DNA and initiates transcription, 5'UTR, gene body containing exons and introns, plus 3'UTR regions. Gene expression is controlled by both proximal and distal regulatory elements – enhancers and repressor elements (Spieth *et al.* 2013). There are two features that differ *C. elegans* from most animals - *trans*-splicing, in which a short RNA sequence, the spliced leader (SL), is spliced onto the 5′ends of mRNAs (Krause & Hirsh 1987), and operon expression– where multiple genes are controlled by a single promoter (Spieth *et al.* 1993). There are 20541 protein-coding genes, which can produce 30446 annotated transcripts. *C. elegans* genome is very compact - exons covers around 29%, and introns around 34% of genome. For comparison, in human genome all genes and regulatory sequences cover only around 3% of total length of around 3 billion base pairs.

1.3 Chromatin – histone modifications and transcription factors

In animals, differences in chromatin are associated with gene expression differences, with chromatin features being broadly similar across animals (Ho *et al.* 2014). For example, open chromatin characterizes actively transcribed regions; H3K4me3 is

associated with active promoters and transcription start sites (TSS); H3K27me3 and H3K9me3 are repressive marks associated with silent developmental genes and constitutive heterochromatin respectively (Bernstein *et al.* 2012); and the H2A.Z histone variant (its *C. elegans* homolog is HTZ-1) is associated with expression in a context-dependent manner (Hu *et al.* 2013b), which means it can be either active or repressive, depending on other proximal histone modifications. Chromatin modifications can be influenced by DNA sequence, but are not necessarily dependent on it, as they can be inherited independently of DNA (Hackett *et al.* 2013). Also, the chromatin landscape changes during organism development, cell differentiation and under the influence of environmental stimuli (Johnson *et al.* 2006).

Transcription factors (TFs) are proteins that influence transcription by binding to promoters, enhancers and silencer regions. The main factor influencing the binding is DNA sequence, which in many of these regions exhibits a defined nucleotide composition often called “the motif”. However, binding is also influenced by chromatin state and many factors will only bind when DNA is accessible (e.g. nucleosome depleted).

Chromatin, TF binding and gene expression (Pol II activity) influence each other. The TFs might be attracted by certain histone modifications and open chromatin, but TF binding can also protect regions against histone deposition and it might maintain an open chromatin state. RNA polymerase binding is speculated to change chromatin conformation, which leads to an increase in active factor binding and higher activity of chromatin remodelers and histone modification writers/erasers (Bernstein *et al.* 2012; Thurman *et al.* 2012). Furthermore, chromatin proteins and modifications interact genetically – a change in expression of certain factors and remodelers might lead to alterations in chromatin state and TF expression. Yet another important factor is histone

modification cross-talk – the process in which one covalent modification influences the function or occupancy of another (Göndör & Ohlsson 2009; Ooi & Wood 2007).

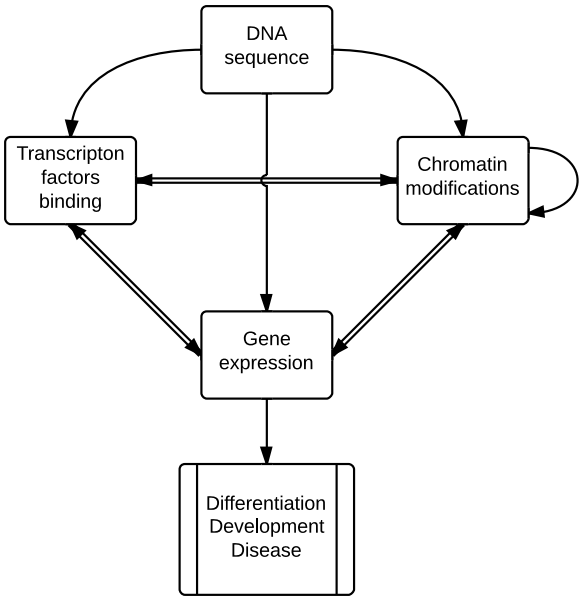


Figure 1 Diagram of proposed regulation model, showing all possible interactions between the gene expression and some regulatory factors measured by ChIP-seq, i.e. chromatin modifications and TF binding.

The literature suggests that there is a vast regulatory network of interactions between DNA, TFs, chromatin modifications and gene expression (Figure 1). The number of TFs in mammalian systems is estimated to be >2000 (Brivanlou & Darnell 2002), and there are potentially hundreds of biologically meaningful histone modifications. Only a small fraction of these factors has been well characterized. In most cases we do not know their molecular function. Binding patterns and co-occurrence with characterised factors are used to infer clues about possible functions of uncharacterised factors. Furthermore, we do not understand the global network of interactions.

1.4 Three-dimensional genome organisation

A further important factor that influences gene expression is 3D chromatin organization. Experimental techniques including chromosome conformation capture (3C) and fluorescence *in situ* hybridization (FISH) have revealed differences in the structural properties and spatial organization of chromosomes. For example, the study of chromosome conformation capture followed by DNA sequencing (HiC) led to identification of topologically associated domains (TADs). TADs are small-scale structural units defined as regions within which genomic interactions appear to be constrained - they are characterized by large number of interactions within the domain, but very few interactions between domains. They correlate with regions of the genome that constrain the spread of heterochromatin, i.e. usually the whole domain can be attributed as active or inactive. The positions of TADs are shared between tissues and can even be conserved across species (Dixon *et al.* 2012). In different organisms, it has been reported that boundaries between TADs are enriched with insulator protein CTCF, transfer RNAs and short interspersed element (SINE) retrotransposons (Dixon *et al.* 2012). However, it is unknown how TADs form and function, and there are numerous models to explain this e.g. cohesin handcuff model (Sofueva & Hadjur 2012). There is also a controversy in the field whether 3C method really captures physical proximity between genomic regions, or is prone to produce false positive interactions between open chromatin due to fixation. Most concerns come from reported poor overlap between proximity ligation methods and FISH (Williamson *et al.* 2014). Regardless of the concerns, TADs do represent an element of genomic organization and are important phenomena to study in order to understand genome function.

1.5 Promoters, enhancers and HOT regions

HOT (high occupancy target) regions are a unique class of regulatory element, as in ChIP-seq assays HOT regions show binding of most known TFs, which cannot be explained by sequence motifs. Genes in proximity to these regions are the very highly transcribed (usually highest decile of expression), ubiquitously expressed, housekeeping ones. In our recent study, we analysed HOT regions in worm and human (Chen *et al.* 2014a), and found that extreme HOT regions are promoters with high levels of CpG dinucleotides, H3K4 trimethylation, accessible chromatin, and they are bound by the non-methyl CpG binding protein CFP-1 - a member of the SET-2/Set1/COMPASS complex. SET-2/Set1/COMPASS complex facilitates deposition of H3K4 trimethylation, and in *C. elegans* contains CFP-1, SET-2, SWD-2.1, ASH-2, DPY-30, WDR-5.1 and RBBP-5 (Chen *et al.* 2014b; Xiao *et al.* 2011). Discovering CpG rich regions recognized by CFP-1 in worms was surprising, since *C. elegans* lacks DNA methylation and was believed not to have CpG enriched regions (termed CpG islands in mammalian systems). Our further investigations have shown that GCG and CGC tri-nucleotides are even better footprints for promoters. The calculated observed vs. expected ratio (O/E) of tri-nucleotides implies that this enrichment is not driven by GC content or presence of CpG. Furthermore, there was a recent publication postulating the existence of di-nucleotide repeat motifs (DRM) in fly enhancers (Yanez-Cuna *et al.* 2014). Therefore, I would like to investigate the role of other tri-nucleotides and longer oligonucleotides in recognition and activity of regulatory sites.

1.6 Transposons and other repeats in *C. elegans* genome

Repetitive DNA elements (repeats) are patterns of nucleotides that occur in multiple copies through the genome. Initially they described as “junk” or “selfish” DNA, later term coming from their self-replicating properties. Many repeats are of viral origin, being evolutionary remains of past viral integration events. It is estimated that between 50 to 69 percent of human genome are repetitive sequences, with roughly similar repeat content is observed in other mammalian genomes (Haubold & Wiehe 2006; Koning *et al.* 2011). In *C. elegans* repetitive elements cover only 16.22% of the genome, according to Dfam 2.0 annotation. Further, the distribution of repeats differs strikingly between *C. elegans* and *H. sapiens*. In human genome the repeats are mostly located in pericentromeric region, forming a large domain of repetitive sequences exclusively. In worm, repeats are dispersed on the chromosomes with higher density of repeats in chromosomal arms. Also, in *C. elegans* repeats are usually interspaced by non-repetitive DNA. This property makes *C. elegans* a particularly useful model to study repeats using next generation sequencing based methods, because (1) *C. elegans* reference genome assembly has most of individual repeat loci annotated and positioned with high confidence, and (2) when using short reads, re-sequencing techniques, such as ChIP-seq and RNA-seq, there is a high probability that reads coming from repetitive regions will have enough unique sequence to confidently map them to a single locus, hence enabling to study the marking and expression of individual repeats, rather than repeat families.

C. elegans genome contain all major classes of repeats:

- Retrotransposons (class 1 transposable elements) – these mobile elements are first transcribed from DNA to RNA, and then RNA product is reverse transcribed to cDNA and integrated at new loci. Retrotransposons are further classified into:
    - LTR retrotransposons – these retroelements are characterized by presence of long terminal repeats (LTR), flanking the coding region, containing two ORFs (*gag* and *pol*), that encodes a set of proteins sufficient for autonomous transposition. In *C. elegans* genome this class is represented by gypsy/Ty3 retrotransposons (Britten 1995).
    - Long interspersed nuclear elements (LINEs) – a group of autonomous retroelements that usually encode two proteins ORF1 and ORF2, which contains reverse transcriptase (RT) and endonuclease (EN) domains, essential for transposition, preceded by Pol II promoter and followed by 3' UTR region (Marin *et al.* 1998).
    - Short interspersed nuclear elements (SINEs) – this class gathers non-autonomous retroelements that relies on LINEs activity for transposition. However, they did not evolve from autonomous elements and are expressed using Pol III, rather than Pol II. They likely originate from the incidental transpositions of Pol III transcripts, rather than from viral integration events. Prominent example of such sequence in *C. elegans* genome is CELE45 (Oosumi *et al.* 1996).
- DNA transposons (class 2 transposable elements) – group of mobile elements that does not require to be transcribed to RNA. These elements are divided in two sub-classes:
    - Cut-and-paste - these elements are excised from donor loci and directly integrated into acceptor loci. Often, we can find a footprint (short DNA

motif) in old excision site that is characteristic for given family. The example of such transposon in *C. elegans* genome is Tc1 (Rosenzweig *et al.* 1983).

- Copy-and-paste – these elements are copied from donor site, to acceptor sites in process involving a nickase – enzyme producing single strand DNA break. The donor site is than repaired by DNA polymerase, and copied fragment is inserted to a new locus. The most studied such element is Helitron1 (Grabundzija *et al.* 2016; Kapitonov & Jurka 2001).

- Satellite repeats – are formed by the arrays of non-coding, tandemly repeating, DNA. In *C. elegans* genome they are interspersed along the holocentric chromosomes (Naclerio *et al.* 1992).

In both *C. elegans* and mammals the first line of defence against the expression of newly integrated sequences is the piRNA pathway, which provides a way to scan the genome, recognise the expression of non-self sequences and block their expression with two distinct mechanisms:

- Cytoplasmic piRNA pathway – this pathway degrades foreign transcripts using cytoplasmic argonaute proteins.
- Nuclear piRNA pathway – the small RNAs are imported into nucleus, and working with nuclear argonaute proteins they establish heterochromatin-like state in the complementary loci, preventing the expression.

In mammals there is a more permanent mechanism for blocking the expression of repeatable elements involving a Krüppel-associated box domain zinc finger proteins (KRAB-ZFPs) (Ecco et al. 2017). However, KRAB domain proteins are absent in *C. elegans* genome, and it is not known if a similar system involving ZFPs and silencing domain is present.

1.7 Heterochromatin and associated factors

Heterochromatin is defined as tightly packed, condensed DNA. Since it is less accessible to Pol II and transcription factors it plays a role in controlling the expression of genes and other DNA elements, for example genomic repeats. In *C. elegans*, heterochromatin is characterised by presence of inactive histone marks, such as H3K9 methylation and localisation of heterochromatin factors, such as HPL-2 (homolog of mammalian HP1). In most animals it flanks centromeres and telomeres. *C. elegans* genome organization provides a unique model to study heterochromatin and its impact on expression – due to holocentric chromosomes with dispersed centromeres heterochromatin domains are scattered around chromosomes, making them easier to study using ChIP-seq and RNA-seq experiments. Further, in *C. elegans* it is possible to remove all traceable H3K9 methylation by knocking out their writers – SET-25 and MET-2, without the introduction of chromosome segregation defects, like in higher animals. This provides the great opportunity to study the impact of H3K9 methylation on expression.

1.8 Genes important for regulation and function of chromatin

The table below summarises the most important genes I characterised in this dissertation. It also gives the mammalian homologs when such are known.

Name	Homolog	Description
hpl-2	HP1	*C. elegans* homologue of heterochromatin protein 1, characterised to have a function in gene regulation and heterochromatin packaging. HPL-2 is important for silencing of transgenes in the germline, germline development, vulval cell fate specification and negatively regulates RNA-mediated interference (RNAi). It contains an N-terminal chromo domain (CD, responsible for binding to methylated H3K9), variable hinge region and a C-terminal- chromo shadow domain (CSD, required for dimerization and protein–protein interactions) (Coustham *et al.* 2006; Eissenberg 2001; Meister *et al.* 2011).
lin-418	Mi-2/ CHD3	Chromatin remodeler, part of a NuRD (Nucleosome Remodeling and Deacetylase) complex, has a broad role in development. It is important for transcriptional repression during development, DNA repair and cell cycle progression (Käser-Pébernard et al. 2014). It plays a role in patterning and germ line development – in *C. elegans let-418* mutant shows ectopic expression of germline genes.
lin-13	N/A	A member of the LIN-35 Rb class of genes involved in vulval development and negative regulation of vulval fates. LIN-13 has multiple zinc fingers domains (C2H2 class, can bind to DNA) and a LXCXE retinoblastoma protein-binding motif.
lin-61	L(3)MBT2	Important for transcriptional repression during development, genome stability in the soma and germline and vulval development (class B synthetic multivulval gene*). Contains 4 malignant brain tumor (MBT) repeats.
met-2	SETDB1	Histone methyltransferase required for normal levels of histone H3K36 and H3K9 trimethylation. Plays a role in vulval development (class B synthetic multivulval gene*), negatively regulates vulval cell fate specification.
set-25	N/A	Histone methyltransferase (C-terminal SET domain) required for normal levels of H3 lysine-9 methylation. *set-25* mutant shows abnormal axonal guidance during development and high somatic mutation rate.
cfp-1	CXXC1	Zinc finger (PHD-type and CXXC-type domains) that binds specifically to non-methylated CpG motifs, with a preference of CpGG through its CXXC domain. A component of the SET1/COMPASS complex that is essential for normal levels of H3K4 methylation.
set-2	STE1/ MLL	Histone methyltransferase required for normal levels of H3K4 methylation. Characterised to act in germline development, postembryonic development, negative regulation of lifespan in adult animals and RNA interference. Component of the SET1/COMPASS complex.
sin-3	SIN3 family	Histone deacetylase complex (HDAC) subunit, has a glutamine and asparagine rich domains. Shows no obvious phenotype in RNAi screens. Required for the deposition of H3K9me2 on autosomal and sex asynapsed chromosome pairs during meiosis (Checchi & Engebrecht 2011).
hda-1	RPD3	Histone deacetylase – deacetylates lysine residues on the N-terminal part of the core histones (H2A, H2B, H3 and H4), plays a role in developmental gene transcriptional regulation and cell cycle progression. HDA-1 is required for embryonic viability, gonadogenesis, endoderm determination and vulval development. Forms large, multiprotein complexes.

Table 1 Summary of genes that were characterised in this dissertation. **Class B synMuv gene* - loss of activity of such gene in the background of a class A or class C synMuv mutant results in a multivulval phenotype, indicating that such gene acts to negatively regulate vulval cell fate specification.

1.9 Chromatin associated protein complexes

Many of biological processes, such as deposition of histone modifications and chromatin remodelling, require a co-operative function of many proteins, which form complexes. These complexes usually consist of targeting proteins, that recognise certain DNA motifs or chromatin features, effector proteins that possess enzymatically active domains, and structural proteins that bring other components together and regulate the complex activity. The table below lists complexes I characterised in this dissertation.

Name	Members	Function
SET-2/ Set1/ COMPASS	WD-2.1 CFP-1 ASH-2 SET-2 DPY-30 RBBP-5 WDR-5.1	This complex is essential for maintaining global levels of H3K4 methylation - a mark widely associated with regulatory elements. H3K4me3 is enriched near active promoters, while H3K4me2/1 are found at both promoter and enhancer regions. SET-2 has a catalytic activity of histone methyltransferase, while CFP-1 targets the SET1 complex to actively transcribed genes through multivalent interactions with chromatin, including recognition of non-methylated DNA and H3K4me3 (Brown *et al.* 2017b).
SIN-3S	HDA-1 SIN-3 ATHP-1 MRG-1	SIN-3 small histone deacetylase complex deacetylate nucleosomes in the vicinity of SIN-3 targeted promoters, resulting in a repressed chromatin structure. The loss of SIN-3S complex subunits SIN-3 and ATPH-1 leads to common chromosome segregation defects in intestinal cells and loss of fertility (Kumar et al. 2012; Kuzmichev et al. 2002).
MLL/ SET-16	UTX-1 ASH-2 SET-16 PIS-1 DPY-30 RBBP-5 WDR-5.1	MLL complexes are responsible for H3K4 methylation at different subsets of homeotic genes (MLL1 and MLL2) and are required for H3K4me1 deposition and enhancer function (Glaser *et al.* 2006; Hu *et al.* 2013a; Wang *et al.* 2009).
NuRD-like	LET-418 LIN-53 LIN-40 HDA-1	Nucleosome Remodelling and Deacetylase complex shows both ATP-dependent chromatin remodelling and histone deacetylase activities. It is implicated in the synMuv B pathway that negatively regulates specification of vulval cell fate (Passannante *et al.* 2010a).
MEC	LET-418 MEP-1	Regulatory complex, which is active during embryonic and early larval development, contains histone remodeler LET-418 and Krüppel-like protein MEP-1. Prevent ectopic germline gene expression in soma by mediating the repressive effect of sumoylation (Passannante *et al.* 2010a; Wu *et al.* 2012).

Table 2 Summary of protein complexes that contain proteins characterised in this dissertation.

1.10 Supervised and unsupervised machine learning methods

Due to increasing amounts of data and high complexity of biological problems, machine learning becomes a prominent tool for better understanding biological phenomena. Machine learning can be divided into supervised and unsupervised algorithms.

In supervised machine learning one typically starts with a large collection of well annotated data and trains a classifier or regression model to perform same task on new, not annotated data. Examples of such algorithms in biology are microscope image classifiers or models for classifying SNPs. Recent addition to the field is deep-learning, when we use deep artificial neural networks with a convolutional layer, that acts as a feature detector and fully connected layer that performs actual classification or regression.

Second approach is unsupervised machine learning – in this scenario one typically has a large set of data, that is not labelled or characterised. The goal of machine learning algorithm is to uncover the structure of data and provide a way to organize or simplify it. Typical tasks include clustering – grouping data by similarity, dimensionality reduction – simplifying large datasets, so they will be easier to visualise and analyse using downstream methods, anomaly detection – detection entities that are dissimilar from other in the set, and noise modelling – removing unwanted variance introduced by technical limitations of experimental method or random biological events.

1.11 Dimensionality reduction in genomics

Datasets in genomics can be represented as mathematical constructs – vectors and matrixes containing numeric values. The dimensionality of such datasets is often very high – at the single base resolution, whole genome sequencing-based experiment (such as ChIP-seq) can be represented as a vector of around 100 million elements for *C. elegans* and over 3 billion elements for *H. sapiens* genome. Even when applying binning – summarizing at a given genomic interval, for example every 1kb with single number, such vectors still have a length of hundreds of thousands to millions of elements. When constructing experimental matrix, this number of elements will be multiplied by number of experiments. Such high dimensionality datasets pose big challenge for machine learning algorithms, analyses and data interpretation. Hence, it is of interest to reduce the dimensionality of datasets while keeping as much of useful information (real biological variance) as possible.

Dimensionality reduction is a machine learning process of reducing the number of random variables by obtaining a set of principal variables. Most commonly used technique of dimensionality reduction is principal component analyses (PCA). In this technique a set of variables is linearly mapped to a lower-dimensional space in the way that maximize the variance. Components are ordered by amount of variance thy capture - the first component captures the highest proportion of variance and the last component the lowest amount of variance. This technique is particularly useful for data visualization, as it allows to present multidimensional datasets as a two-dimensional scatter plot.

1.12 Factor Analyses, sparsity and random errors in data

Factor analyses (FA) is a machine learning technique related to PCA. While PCA extracts extracting linear combinations of observed variables, FA uses observed variables to construct a formal model that predicts observed data from theoretical unobservable factors. In factor analyses one tests or assumes that observed variables are a consequence of the underlaying latent structure of the data. FA decomposes original data matrix into two matrixes – a latent factor matrix and factor loading matrix. Multiplication of these matrixes produce the reconstruction (approximation) of original data matrix – the difference between original and reconstructed matrix is a model error.

Factor analyses can be extended with two concepts particularly useful for genomic data – sparsity and noise model. Sparsity in genomic experiments, such as ChIP-seq, arises from the fact that vast majority of chromatin associated proteins and modifications occupies only a very small fraction of the genome. This means that most of elements of vector representing such experiment will be near zero values representing experimental noise. In practical terms a sparsity is implemented as binary matrix (all values are either 0 or 1) of the same dimensionality as latent factor matrix. These two matrixes are multiplied element wise, masking out the features (e.g. genomic loci) that are not relevant for the model. Such procedure reduces computational complexity and increases dimensionality reduction performance of factor analyses.

Noise model accounts for random errors and other sources of randomly distributed noise in the data. It is often implemented as a multivariate Gaussian distribution, whose parameters are estimated from the experimental data. Noise model can help removing unwanted variance from the data, increasing computational efficiency and improving data representation. It is of particular importance for ChIP-seq data, where nonspecific antibody binding can introduce a high level of random noise to experiments.

2 HETEROCHROMATIN FACTORS

***Collaboration note**. This chapter contains results and figures published in McMurchy, Stempor, Gaarenstroom et. al, in 2017 on which I was the lead bioiformatician and shared first author. All experimental work was performed by co-authors. In addition, Ni Huang performed the enrichment analysis in Figure 11.*

In this investigation I focused on an interesting biological problem - the role of heterochromatin factors in genome regulation. To study how heterochromatin forms and functions, our lab undertook the study of a set of five genes implicated in heterochromatin function or transcriptional repression: HPL-2/HP1, zinc finger protein LIN-13, MBT-repeat protein LIN-61, LET-418/Mi-2 and H3K9me2 histone methyltransferase MET-2/SETDB1, as well as histone modifications H3K9me2 and H3K9me3, which are the canonical heterochromatin marks. In addition, I analysed genes and repeats expression profiles in chromatin factor and small RNA pathway mutant backgrounds: *hpl-2*, *let-418ts*, *lin-61*, *hpl-2::lin-61*, *set-25::met-2*, and proteins connected to 22-G and 21-U RNA pathways: *nrde-2*, *nrde-2::let-418ts*, *prg-1.* For this project I have generated pipelines for ChIP-seq and differential gene expression analyses, which capture specific requirements of analysing repetitive sequences.

We used ChIP-seq technique to map binding locations of proteins we anticipated to have a role in heterochromatin formation and maintenance. I performed initial analyses of binding patterns, and it revealed that heterochromatin factors co-localise with repetitive elements and might have a role in their regulation. This observation re-shaped our initial question into a more specific one: how heterochromatin factors, with help of small RNA pathways, regulate repetitive elements expression and prevent germ line stress.

From previous studies we know that repetitive elements, derived from transposons, must be silenced to protect genome integrity (Scheifele *et al.* n.d.; Ward *et al.* 2013). Heterochromatic proteins have been implicated in silencing repeats (Saksouk *et al.* 2015), however exact mechanisms remain poorly understood. Also, *C elegans* is a great model to study H3K9me3 and heterochromatin factors involvement in expression regulation, because due to holocentric centromeres it does not exhibit severe chromosome segregation defects due to a perturbation of centromeric heterochromatin, like most of other higher animals (Kellum & Alberts 1995). Mutants of heterochromatin genes have similar phenotypes. They have abnormal germ line development, reduced fertility, and increased germ line apoptosis. Other shared phenotypes are slow growth, DNA repair defects, and increased germ line mutations. Several genes have been shown to be needed for piRNA pathway silencing (Ceol *et al.* 2006; Koester-Eiserfunke & Fischle 2011; Meléndez & Greenwald 2000; Schott *et al.* 2006; Thomas *et al.* 2003; von Zelewsky *et al.* 2000). Furthermore, genetic interactions between these genes indicate that they function together (McMurchy *et al.* 2017).

In summary, to understand the relationships and functions of heterochromatin factors, I used computational methods to analyse profiles of chromatin factor binding. We employed ChIP-seq to map their binding locations in young adults and RNA-seq of

mutants to determine their roles in gene and repeat expression. ChIP-seq of H3K9me2 and H3K9me3 was also performed for comparison to heterochromatin protein profiles. Colleagues in the lab performed the experimental work and I performed most of the computational analyses. The ChIP-seq data used for this project are summarized in Table 3, while the replicates are shown in Table 4.

ID	Factor	Antibody	LotNumber	Stage	Strain	Crosslinker
AA560	H3K9me2	302-32369	11005	YA	N2	F
AA561	H3K9me2	302-32369	11005	YA	N2	F
AA601	H3K9me3	ab8898	339901	YA	N2	F
AA602	H3K9me3	ab8898	339901	YA	N2	F
AA249	HPL2	Q2324	Q2324	YA	N2	E
AA250	HPL2	Q2324	Q2324	YA	N2	E
AA568	LET418	Q3861	Q3861	YA	N2	E
AA569	LET418	Q3861	Q3861	YA	N2	E
AA566	LIN13	Q0838	Q0838	YA	N2	E
AA567	LIN13	Q0838	Q0838	YA	N2	E
AA564	LIN61	Q3520	Q3520	YA	N2	E
AA565	LIN61	Q3520	Q3520	YA	N2	E
AA570	MET2	Q2943	Q2943	YA	N2	E
AA571	MET2	Q2943	Q2943	YA	N2	E

Table 3 Summary of ChIP-seq experiments used in the study of heterochromatin factor binding locations. F denotes formaldehyde, while E denote EGS (ethylene glycol bis(succinimidyl succinate)) crosslinkers.

Factor	Antibody	Strain	Stage	Experiments	Correlation [100bp bins]
LIN61	Q3520	N2	YA	AA564|AA565	0.9301
LIN13	Q0838	N2	YA	AA566|AA567	0.9528
H3K9me2	302-32369	N2	YA	AA561|AA560	0.9840
HPL2	Q2324	N2	YA	AA250|AA249	0.9715
LET418	Q3861	N2	YA	AA568|AA569	0.9553
MET2	Q2943	N2	YA	AA570|AA571	0.9133
H3K9me3	ab8898	N2	YA	AA601|AA602	0.9758

Table 4 The summary of ChIP-seq experiment replicates used in the study of heterochromatin factor binding locations. Correlation column denotes Pearson correlation coefficient between rBEADS normalised tracks in the replicate in 100bp bin resolution.

2.1 Heterochromatin factors are well correlated with each other and with H3K9me2, but not with H3K9me3

First, I analysed the binding of 5 heterochromatin (HC) factors using ChIP-seq data. To establish a chromatin context in which five investigated factors bind I further analysed ChIP-seq data of H3K9me2 and H3K9me3. H3K9 methylation is facilitated by MET-2 and SET-25 methyltransferases, and HPL-2/HP1 was reported to directly bind H3K9me (Eskeland *et al.* 2007). I wanted to establish, if in *C. elegans* H3K9me is also co-localising with heterochromatin factors, and if so which level of methylation (me2 vs. me3) is dominant on these marks.

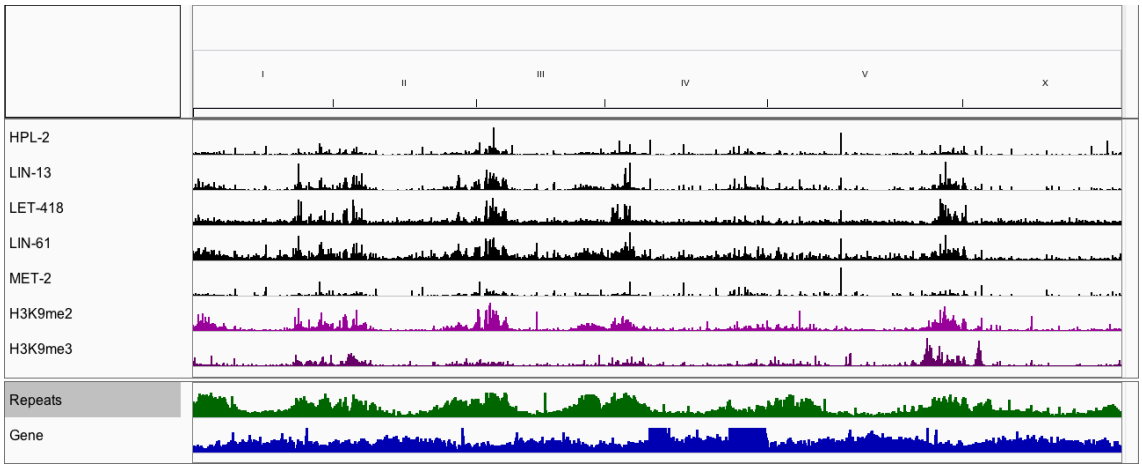


Figure 2 Heterochromatin factors are enriched on the distal arms of autosomal chromosomes. Visualization in IGV genome browser. HPL-2, LIN-13, LIN-61, LET-418, and MET-2 are shown in black, H3K9me in purple and repeats and gene density in green and blue, respectively.

On the global scale all factors are enriched on the distal arms of autosomal chromosomes (**Figure 2**). This pattern was observed before for H3K9me3, and HPL-2 (Garrigues *et al.* 2015; Liu *et al.* 2011b; Sha *et al.* 2010). I have also observed similarities in patterns of these proteins at specific loci. Further analyses revealed that this similarity is genome-wide – the correlation analyses have shown a significant, positive correlation between all 5 sets (**Figure 3**). Furthermore, the signals correlated very well with H3K9me2, but not H3K9me3.

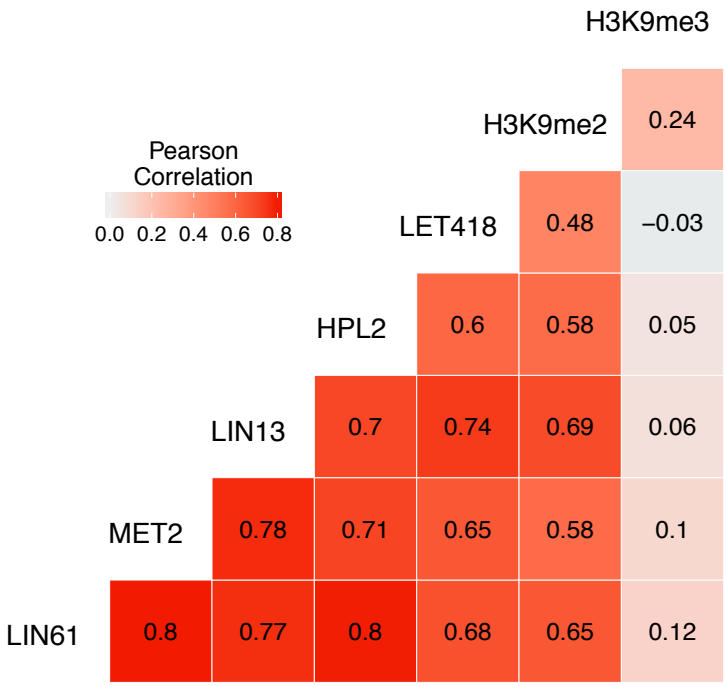


Figure 3 Heatmap showing Pearson correlation coefficients for ChIP-seq track combined replicates. Tracks were sampled in 100bp windows.

Further, I observed that the genomic distributions of H3K9me2 and H3K9me3 differ. H3K9me3 forms broad domains, while H3K9me2 binds in defined loci, forming sharp peaks similar to transcription factors. Moreover, H3K9me2 co-localises with five HC factors, while H3K9me3 shows no enrichment in overlapping loci (**Figure 4**). It should be noted that there is very little cross-reactivity between H3K9me2 and H3K9me3 antibodies we selected for ChIP-seq experiments (**Appendix 8.3**).

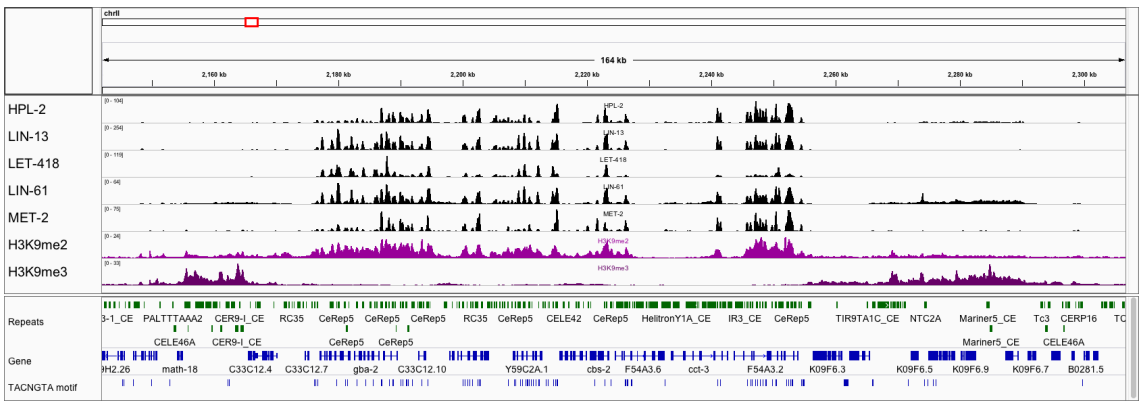


Figure 4 H3K9me2 co-localises with five HC factors, while H3K9me3 shows no enrichment in overlapping loci. Visualization in IGV genome browser. HPL-2, LIN-13, LIN-61, LET-418, and MET-2 are shown in black, H3K9me in purple, and repeats and gene density in green and blue, respectively.

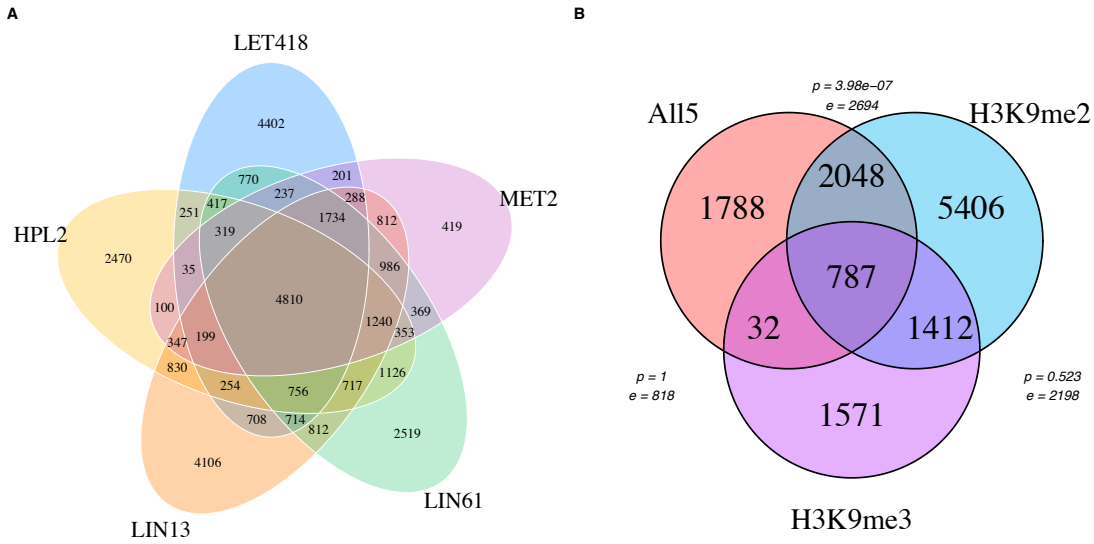


Figure 5 Heterochromatin factors overlap well with each other, and *All5* set overlaps significantly with H3K9me2, but not H3K9me3. (A) Five-way Venn diagram representing overlaps between 5 HC factors. I called 4810 intersections set of all factors *"All5"*. (B) This set significantly overlaps with H3K9me2, but there is no significant overlap with H3K9me3. P-values (p) and expected values (e) calculated using Fisher test and hypergeometric distribution.

Next, I wanted to narrow my investigation only to regions bound by one or more chromatin factors. I identified peak regions for each dataset. LIN-13 was enriched in the highest (19313) and MET-2 in the lowest number (12449) of loci (**Table 5**). I found a total of 33301 loci enriched for at least one factor, with majority (58%) bound by more than 1 factor. The biggest peak group, with 4810 sites, is where all five factors bind (**Figure 5**). Sites uniquely bound by only one factor are rare, ranging from 3.4% of unique peaks for MET-2 to 27.4% for LET-418. To facilitate further combined factor analyses I created two sets: (1) "*All5*" for loci bound by all factors and (2) "*Any5*" for loci bound by at least one factor (**Figure 6**, **Table 5**).

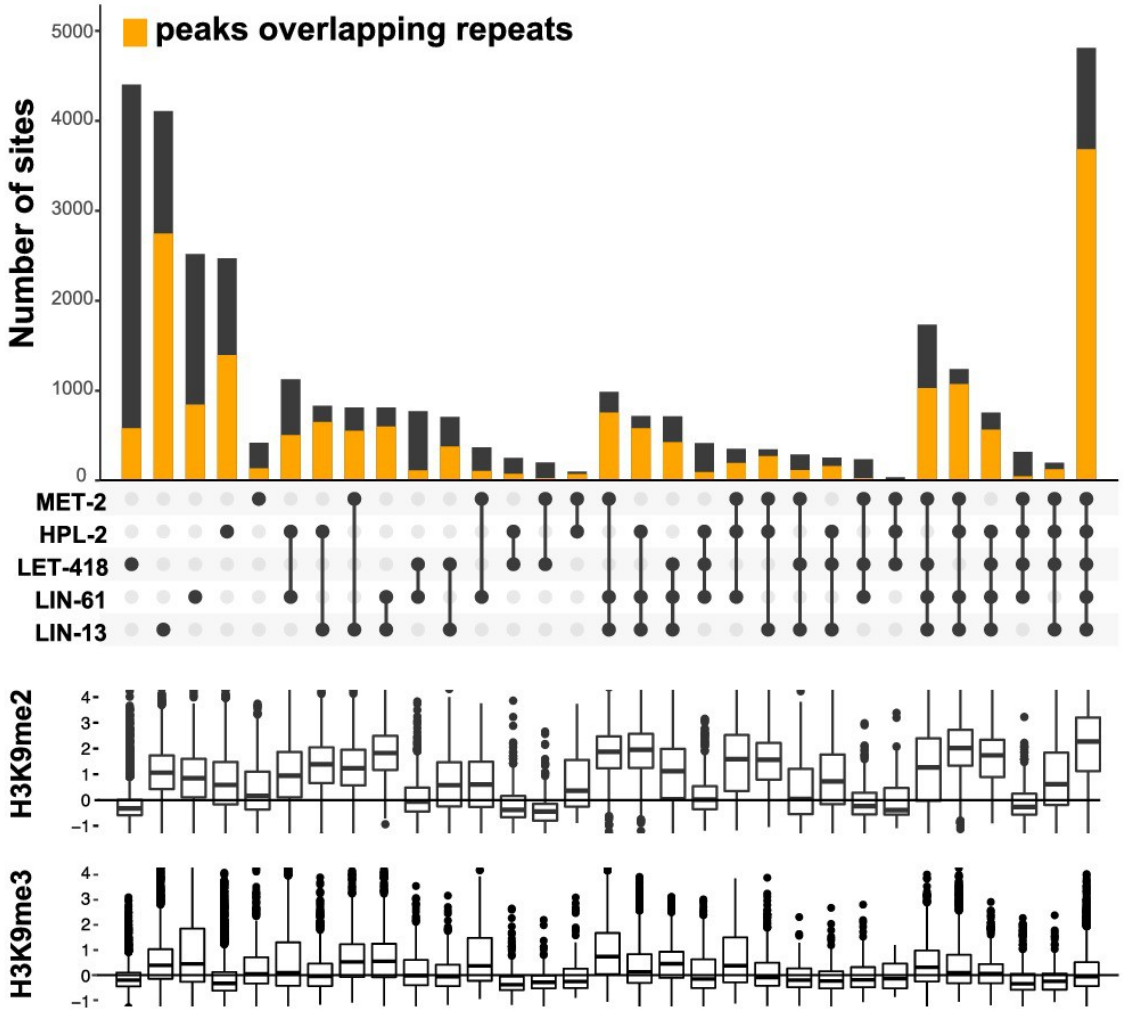


Figure 6 UpSet plot showing overlaps between HC factors and the 33,301 *Any5* peak calls. Dots indicate which peak calls are assessed, and dots connected with lines indicate overlap groups. Bars show total number of peaks per class, the orange portion denoting overlap with repeats. In the bottom the boxplots show levels of H3K9me2 and H3K9me3. The positions that overlap all five factors constitute the largest class (n = 4810).

Dataset	Total Number of Peaks	Number (%) unique to dataset	Number (%) overlapping repeats
HPL-2	14224	2470 (17.4 %)	9497 (66.8%)
LET-418	16095	4402 (27.4%)	7452 (46.3%)
LIN-13	19313	4106 (21.3%)	13708 (71.0%)
LIN-61	17879	2519 (14.1%)	10634 (59.5%)
MET-2	12449	419 (3.4%)	8219 (66.0%)
Any5	33301	NA	17932 (53.8%)
All5	4810	NA	3683 (76.6%)

Factor	Number (%) of repeats with factor bound
HPL-2	17026 (27.3%)
LET-418	15931 (25.6%)
LIN-13	22213 (35.6%)
LIN-61	20026 (32.1%)
MET-2	13779 (22.1%)
Any	28974 (46.5%)
All	8002 (12.8%)

Table 5 Number of peaks called for heterochromatin factor ChIP-seq experiments and the numbers of regions in *All5* and *Any5* sets, plus the number of peaks unique to each dataset and the number of peaks overlapping repetitive elements. Last column denotes number of Dfam 2.0 repetitive elements (n = 62331) that have factors bound based on IDR peak calls overlap.

In agreement with observation for signal distribution, peak loci are enriched on distal chromosomal regions of autosomes. Most peaks localise to intergenic or intronic regions. Intergenic peaks are more present in central chromosome regions, while intronic peaks localise to distal chromosome arms (**Figure 7**).

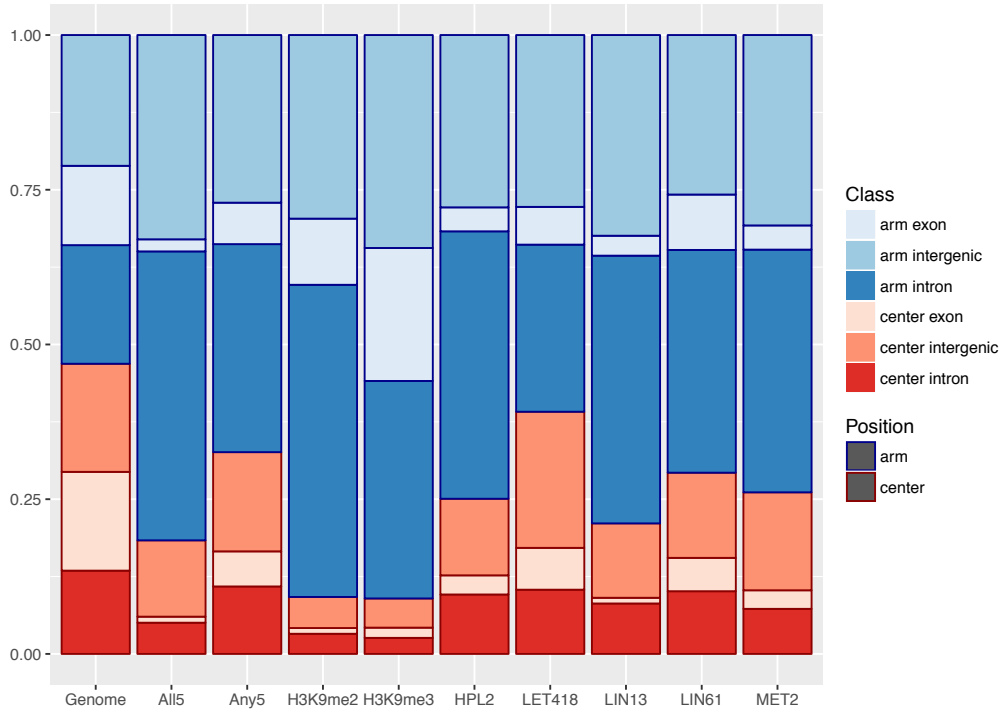


Figure 7 The genomic distribution of heterochromatin factors. The blue outline and shades of blue denotes arm, and red outline and shades of red denote central chromosome region. The first bar "Genome" shows what fraction of *C. elegans* genome these annotations constitute.

The enrichment of H3K9me2 and H3K9me3 varies between the loci, with *All5* class showing the highest enrichment for H3K9me2, but no enrichment for H3K9me3 (**Figure 6**). To determine if the enrichment of H3K9me2, and no enrichment of H3K9me3 at 5 HC factors binding loci is a genome wide feature, I quantified H3K9me2 and H3K9me3 signal on 33301 *Any5* sites (**Figure 8**). Indeed, H3K9me2 shows strong enrichment at these sites, with median rBEADS score ~2, denoting twice or more enrichment over noise level for H3K9me2. H3K9me3 median rBEADS signal equals 1, meaning no enrichment over noise level. I conclude that HPL-2, LET-418, LIN-13, LIN-61, and MET-2 loci extensively overlap with each other and H3K9me2 marks.

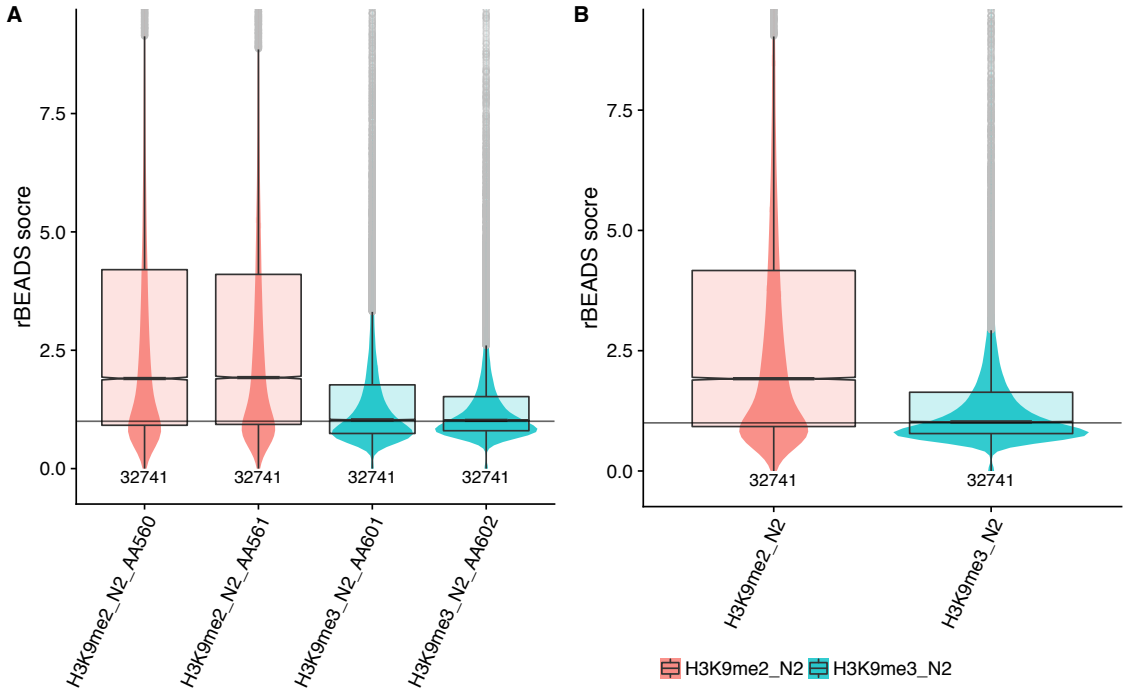


Figure 8 Heterochromatin factors show strong enrichment of H3K9me2, but no enrichment for H3K9me3. The quantification of rBEADS normalised signal at *Any5* sites. (A) Individual replicates, and (B) combined replicates.

2.2 Heterochromatin factors are enriched on repetitive elements

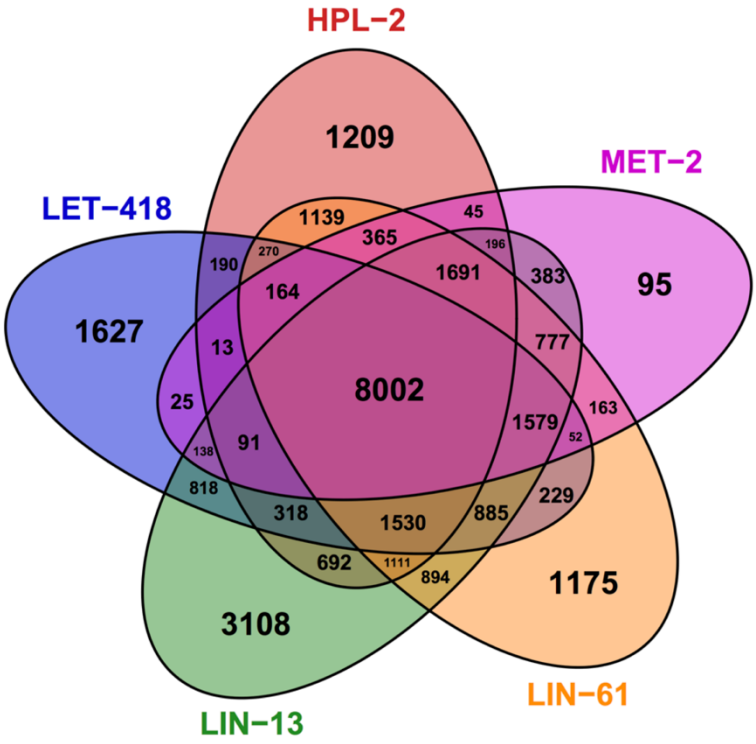


Figure 9 Venn diagram showing overlap in repeats marked by five chromatin factors. The biggest group, 8002 contains repeats bound by all factors. The most numerous group bound by single HC factor is LIN-13 – it was expected, since LIN-13 is a zinc finger protein directly binding to DNA. On the other hand, MET-2 has almost no unique binding, in vast majority of the cases being accompanied by at least one other factor.

It was reported previously that HPL-2 localises to repetitive elements in *C. elegans* embryo (Garrigues *et al.* 2015). I investigated if this is true for young adult *C. elegans* and if other heterochromatin factors bind repetitive regions. I used DFam 2.0 (Hubley *et al.* 2016) annotation, which classifies 62,331 individual repetitive elements into 184 families. The families can be grouped into three types: DNA transposon, retrotransposon and satellite/unknown. I observed that heterochromatin factors are generally enriched on repetitive elements, with LIN-13 showing the strongest enrichment (71% of peaks bind to repeats) and LET-418 showing the weakest enrichment of 46.3%. 53.8% of all HC loci (Any5 set) overlap with repeats. There is even stronger enrichment for *All5* set, with 76.6% sites overlapping repeats. Out of total

62331 annotated repeats, 46% overlap *Any5* set and 8002 (13%) *All5* set (Figure 9 and Table 5).

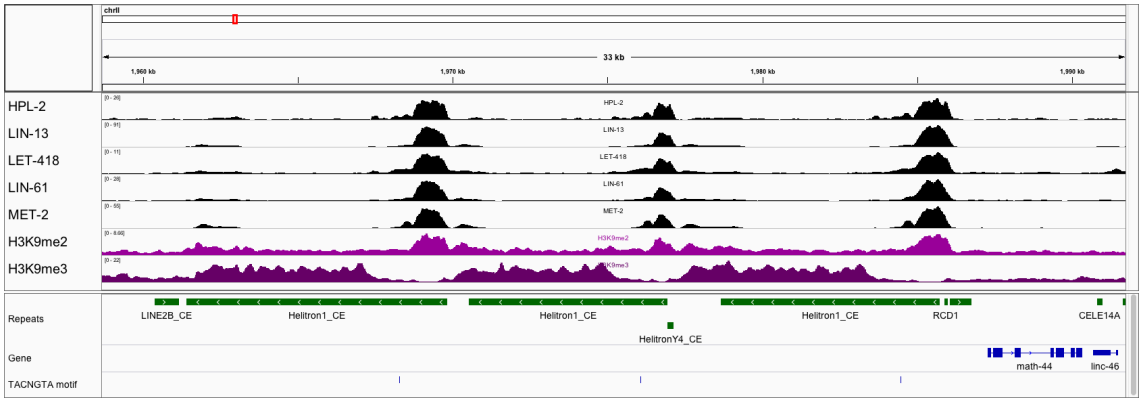


Figure 10 Heterochromatin factors (HPL-2, LIN-13, LIN-61, MET-2, and H3K9me2) have a very strong association with Helitron1. IGV screenshot showing the Heliron1 cluster on ChrII.

Heterochromatin factors overlap with both DNA transposon, retrotransposon, as well as with satellite/unknown repeats – total of 180 out of 184 families are associated with a heterochromatin factor peak, and 105 repeat families with two or more factors. I also noticed that most of the factors (HPL-2, LIN-13, LIN-61, MET-2, and H3K9me2) have a very strong association with Helitron families, in particular with Helitron1 (**Figure 10**) and Helitron2. LET-418 is an outlier – it shows lower enrichment on repeats than four other factors (**Figure 6**).

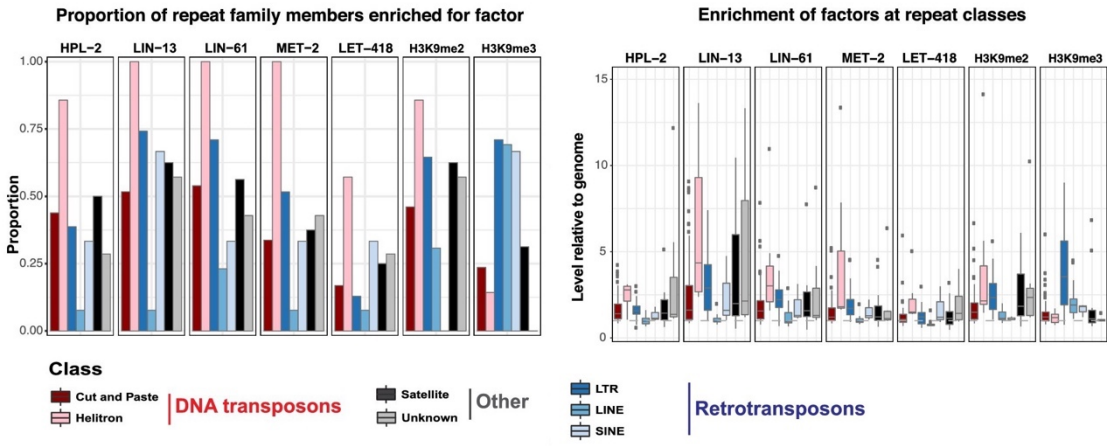


Figure 11 H3K9me2 and H3K9me3 are enriched on different classes of repeats. Left panel: proportion of repeats in the family bound/marked by individual factors. Right panel: relative enrichment of factors at repeat classes. Enrichment analyses were performed by Ni Huang.

H3K9me2 and H3K9me3 are enriched on different classes of repeats: H3K9me2 is more associated with DNA transposon and satellite/unknown repeats, while H3K9me3 associated with retrotransposon – particularly with LINE and SINE elements (**Figure 11**). Therefore, I observed that all five heterochromatin factors are enriched on repetitive elements – this suggests that these factors might have a collaborative role in supressing the expression from repetitive DNA.

2.3 H3K9me2 and heterochromatin factors are enriched on telomeres

In the next step I checked if five heterochromatin factors are enriched on telomeric repeats. To do this, I developed an assay for analysing telomere enrichment from ChIP-seq data by counting reads with the repetitive "GCCTAA" motif. I defined telomeric reads as those having 5 or 6 "GCCTAA" exact matches of motifs in trimmed, 36bp read. I determined background levels of motif enrichment using 129 inputs. Then I calculated the depth of reads coverage normalised fold enrichment over mean of inputs. To assess the statistical significance of enrichment I used one sided Mann–Whitney U test (2 replicates for each factor vs. input background of 129 experiments).

I found that all 5 chromatin factors are significantly enriched on telomeres (p-values < 0.05), with LET-418 showing the lowest enrichment of 3.47-fold. It should be noted that two factors, LIN-13 and MET-2, are very strongly enriched with 11.33-fold and 14.28-fold enrichments and p-value < 0.01. H3K9me2 shows even stronger enrichment of 20.02 (p-value < 0.01). However, H3K9me3 is not enriched at all showing not significant (p-value ~1) depletion comparing to input (**Table 6**). This result suggests that heterochromatin factors in collaboration with H3K9me2 regulate the telomeric regions, but H3K9me3 is not required for this process.

Info	Telomeric enrichment normalised to pooled inputs

Factor	Id	Individual Replicates	Combined replicates	U-test p-value
H3K9me2	AA560 AA561	22.33 17.7	20.02	0.0079
H3K9me3	AA562 AA563	0.9 0.86	0.88	~1.0000
HPL2	AA249 AA250	2.1 6.27	4.18	0.0190
LET418	AA568 AA569	3.5 3.43	3.47	0.0131
LIN13	AA566 AA567	12.37 10.3	11.33	0.0079
LIN61	AA564 AA565	5.77 4.77	5.27	0.0107
MET2	AA570 AA571	16.2 12.37	14.28	0.0079

Table 6 Heterochromatin factors and H3K9me2, but not H3K9me3 are enriched on telomers. Enrichment of telomeric binding for indicated factors was assayed using ChIP-seq, p-values come from Mann-Whitney U-test. Reads were determined as telomeric, if they contain at least five perfect matches to the telomeric motif (GCCTAA sequence) in 36bp read (max 6 repeats for this read length). All experiments were performed in young adult developmental stage.

2.4 HPL-2, LIN-13, LIN-61, LET-418, and MET-2 are required for repetitive DNA silencing

After determining that all five factors are localising to repetative regions I wanted to determine if they are required for transcriptional silencing of these regions. The silencing of repeats is vital for genomic stability and proper germline function. To validate this possibility, I analysed the RNA-seq expression profiling experiments for mutants of all five chromatin factors and wild type (N2) strain.

The exception here is experiment for MET-2, which is the writer of H3K9me1 and H3K9me2 histone modification. However, all three levels of histone 3 lysine 9 methylation are still present in *met-2*, at lower levels than in wild type. This is due to partially redundant action of a different histone methyltransferase - SET-25 (Towbin *et al.* 2012). For this reason, we analysed *met-2 set-25* double mutants, in which H3K9

methylation is undetectable. It should be noted that there is a third divergent methyltransferase SET-32 implicated in H3 lysine-9 methylation in germline.

Strain	Growth Condition	ID	Aligned Reads	Aligned to gene models	Aligned to repeat model
hpl-2	20C	rAM061	21.61	16.15	0.61
hpl-2	20C	rAM062	22.48	16.18	0.59
hpl-2::lin-61	15C	rAM076	24.6	16.48	0.68
hpl-2::lin-61	15C	rAM084	18.61	13.67	0.89
let-418ts	20C	rAM059	21.85	15.04	0.59
let-418ts	20C	rAM060	22.26	15.63	0.6
lin-13	15C	rAM070	22.42	15.49	1.62
lin-13	15C	rAM069	25.13	17.86	0.65
lin-61	20C	rAM063	22.84	15.98	0.55
lin-61	20C	rAM064	24.35	17.86	0.61
N2	20C	rAM057	21.97	15.16	0.54
N2	20C	rAM058	21.82	15.81	0.57
N2	15C	rAM077	20.44	14.57	0.48
N2	15C	rAM078	22.7	15.76	0.73
nrde-2	15C	rAM079	22.13	16.72	0.57
nrde-2	15C	rAM080	34.13	23.95	0.9
nrde-2::let-418ts	15C	rAM081	22.98	15.98	0.64
nrd-e2::le-t418ts	15C	rAM082	24.13	16.54	0.67
prg-1	15C	rAM075	25.25	17.95	0.67
prg-1	15C	rAM083	27.74	19.99	1.55
set25-met2	20C	rAM087	35.5	23.59	2.29
set25-met2	20C	rAM088	36.9	25.15	1.76

Table 7 Summary of RNA-seq samples analysed in heterochromatin factors study. All experiments were performed in young adult developmental stage using total RNA cell fraction (cytoplasmic and nuclear RNA was profiled). Aligned reads are given in millions.

I wanted to determine whether individual repeats were de-repressed and whether families that these repeats were members of were also affected. Alternatively, only the individual repeats might be de-repressed, not the whole families. Analysing repeat expression is not a trivial problem, since by definition their DNA sequence is repeated multiple times in the genome, causing reads to multi-map, i.e. reads aligner finds more than one optimal position for the read in the genome. Also, repeat expression is in general much weaker than gene expression – one must have to be extra careful not to overestimate the significance of weak effects. Finally, in *C. elegans* genomes repeats

often reside in close proximity of coding genes or in introns. This can be viewed as both advantage and disadvantage for this analysis. Repeats are often bordering with unique DNA sequence, allowing me to find a read that covers both repeated and unique regions, to profile repeat expression and map it to an individual loci. However, close proximity of genes means that repeats expression can be driven by surrounding genes. In order to solve above problems, I created a procedure that determines if individual repeat is truly differentially expressed by integrating two separate analyses of repeat expression – based on all and non-multimapping reads, with coding genes differential expression (see method sections 6.1.4 and 6.1.5 for the detailed algorithm).

In brief, my pipeline starts with determining differentially expressed genes. Then, differentially expressed (DE) repeats are found based on all mapped reads. Repeats that reside in introns, in proximal distance to TSS and translation termination sites (TTS), or in operons of DE genes are discarded from analyses. Then the same procedure is repeated on non-multimapping, high quality reads (mapping quality over 10). Repeats found differentially expressed in both procedures are flagged as validated. For fold change and false discovery ratio (FDR, i.e. scaled p-value) I use estimates derived from all reads.

My collaborators collected samples in chromatin factors and small RNA pathways mutant backgrounds: *hpl-2*, *let-418ts*, *lin-61*, *hpl-2::lin-61*, *set-25::met-2*, and proteins connected to 22-G and 21-U RNA pathways: *nrde-2*, *nrde-2::let-418ts*, *prg-1* (**Table 7**). For each strain I analysed two biological replicates and calculated the differential expression (mutant strain vs. wild type) for coding genes annotated in Ensembl, and 62,331 repeat elements annotated in Dfam2.0.

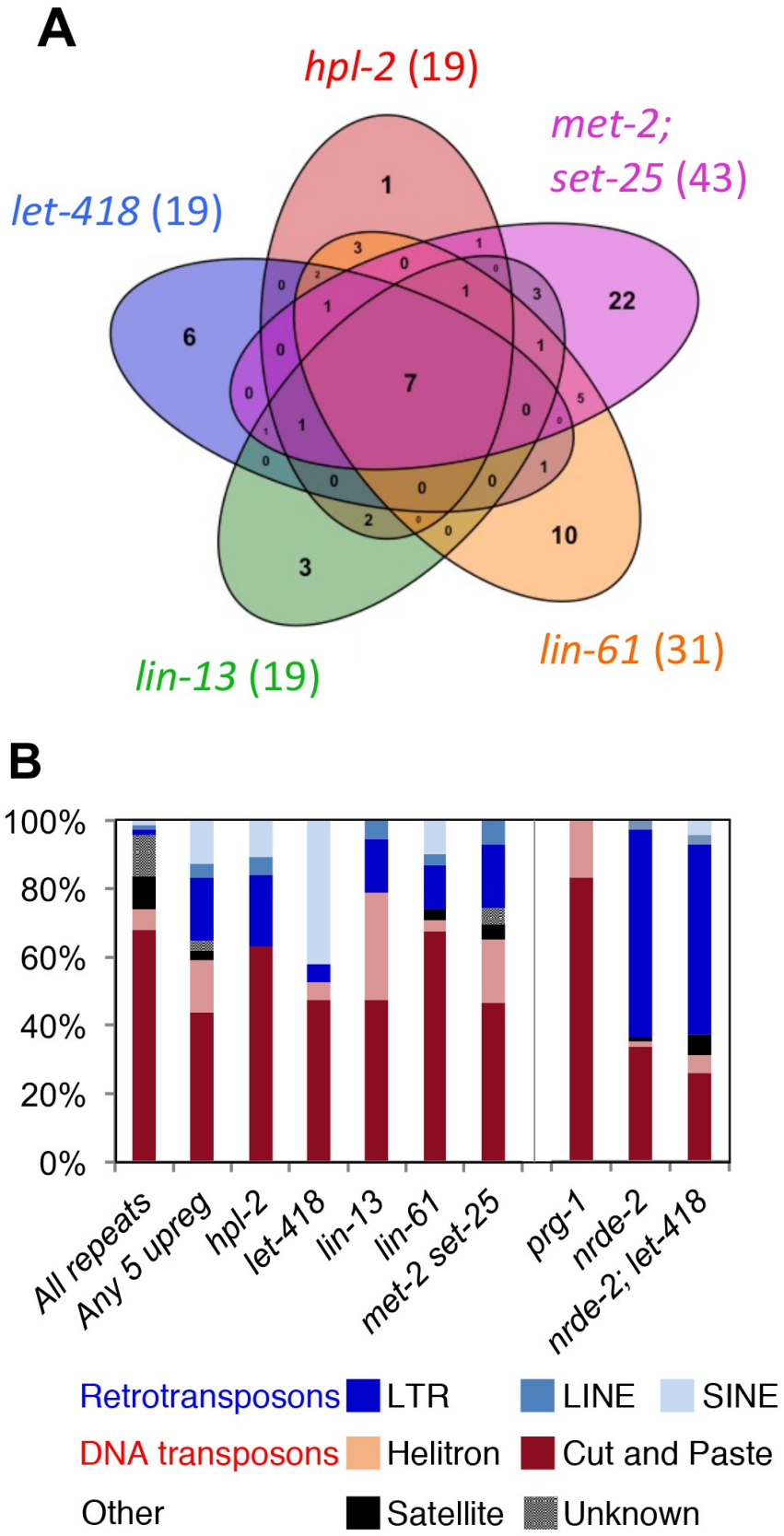


Figure 12 Upregulated repetitive elements. (A) The Venn diagram showing overlap between repeats upregulated in different heterochromatin mutants. (B) The bar pot showing what classes of upregulated repeats are represented in heterochromatin mutants and All5/Any5 sets.

Family	Class	hpl-2	let-418	lin-13	lin-61	met-2 set-25	N strains
MIRAGE1	Cut_and_Paste	11	8	8	11	8	5
CER9-I_CE	LTR	2	1	2	1	4	4
Vingi-1_CE	LINE	1	0	1	1	1	4
CELE45	SINE	2	8	0	3	0	3
CERP2	Cut_and_Paste	0	0	1	1	1	3
Helitron1_CE	Helitron	0	1	6	0	3	3
CEMUDR1	Cut_and_Paste	0	1	0	5	0	2
CEMUDR2	Cut_and_Paste	0	0	0	1	1	2
CER16-2-I_CE	LTR	1	0	0	1	0	2
CER9-LTR_CE	LTR	1	0	1	0	0	2
HelitronY1_CE	Helitron	0	0	0	1	2	2
RC35	Satellite	0	0	0	1	2	2
TC4	Cut_and_Paste	0	0	0	2	2	2
CELE1	Cut_and_Paste	0	0	0	0	1	1
CELE14B	Cut_and_Paste	0	0	0	0	2	1
CELE4	Cut_and_Paste	0	0	0	0	1	1
CELE46B	Cut_and_Paste	1	0	0	0	2	1
CER1	LTR	0	0	0	0	3	1
CER10-I_CE	LTR	0	0	0	0	1	1
CER15-I_CE	LTR	0	0	0	1	0	1
CeRep5	Unknown	0	0	0	0	1	1
CERP3	Unknown	0	0	0	0	1	1
HelitronY1A_CE	Helitron	0	0	0	0	3	1
LTRCER1	LTR	0	0	0	1	0	1
NeSL-1	LINE	0	0	0	0	1	1
PAL8C_4	Cut_and_Paste	0	0	0	0	1	1
PALTTTAAA2	Cut_and_Paste	0	0	0	0	1	1
RTE1	LINE	0	0	0	0	1	1
Turmoil2	Cut_and_Paste	0	0	0	1	0	1

Table 8 Repeat families with at least one member upregulated in *hpl-2*, *let-418*, *lin-13*, *lin-61*, or *met-2 set-25* mutant strains. The mutant strain columns give the number of repeats upregulated in the given family and mutant, and "N strains" column denotes the number of strains upregulating the element in family.

I observed upregulation of repetitive elements in all five mutant strains. However, I found only 71 out of 62,331 individual repeats, representing 29 out of 184 families being upregulated in any of the mutant strains. I validated 61 of these elements to be upregulated when counting only reads uniquely mapping to a single genomic location. Strikingly, even with such low number of overexpressed repeats, I observed a

significant overlap between upregulated repeats, with 41% of repeats being upregulated in more than one mutant strain (**Figure 12A**).

We found seven repeats being upregulated in all strains - all these repeats belonged to MIRAGE1 DNA transposable family (**Table 8**). Other prominent repeats are Vingi-1 and CER9-1 upregulated in 4 mutants, and CERP2, CELE45 and Helitron1 upregulated in 3 mutants (**Table 8**). In each strain the majority of elements are DNA transposons, however *hpl-2* and *let-418* are also enriched for retrotransposons (**Figure 12B** and **Figure 8**). Also, some classes are specific for some mutants, for example Helitrons are particular enriched in *lin-13* (6 out of 19 upregulated repeats are Helitrons), while SINE retrotransposons are upregulated primarily in *let-418* (6 out of 19 upregulated repeats are SINE) (**Figure 12B)**.

Class	Type	hpl-2	let-418	lin-13	lin-61	met-2 set-25	N strains
Cut_and_Paste	DNA Transposon	12	9	9	21	20	5
LINE	Retrotransposon	1	0	1	1	3	4
LTR	Retrotransposon	4	1	3	4	8	4
Helitron	DNA Transposon	0	1	6	1	8	3
SINE	Retrotransposon	2	8	0	3	0	3
Satellite	Undefined	0	0	0	1	2	2
Unknown	Undefined	0	0	0	0	2	1

Table 9 Repeat classes with at least one member upregulated in *hpl-2*, *let-418*, *lin-13*, *lin-61*, or *met-2 set-25* mutant strains. The mutant strains columns give the number of repeats upregulated in the given class and mutant, and "N strains" column denote the number of strains upregulating the element in family.

In further analysis I wanted to ask if upregulation of individual repeats extend to repeat families. For this porpoise I collapsed individual repeats into groups (184) and counted the tags for family, similar as tags on exons constitute gene count. Then I obtained differential expression estimates for families.

Looking at repeats de-repressed in families, I have also found that MIRAGE1 is upregulated in all heterochromatin mutants (**Figure 13**). The other families differ from

individual expression results. This is because this analysis is heavily biased towards small families containing similar repeats. For example, MIRAGE1 family has only 69 members and CeRep57 has 19 in contrast to Helitron1 having 432 members, and HelitronY1A – 1628 members. Hence, I see many upregulated retrotransposon families, especially LTR families, since these families are generally less numerous than satellite or DNA transposons (**Figure 13**).

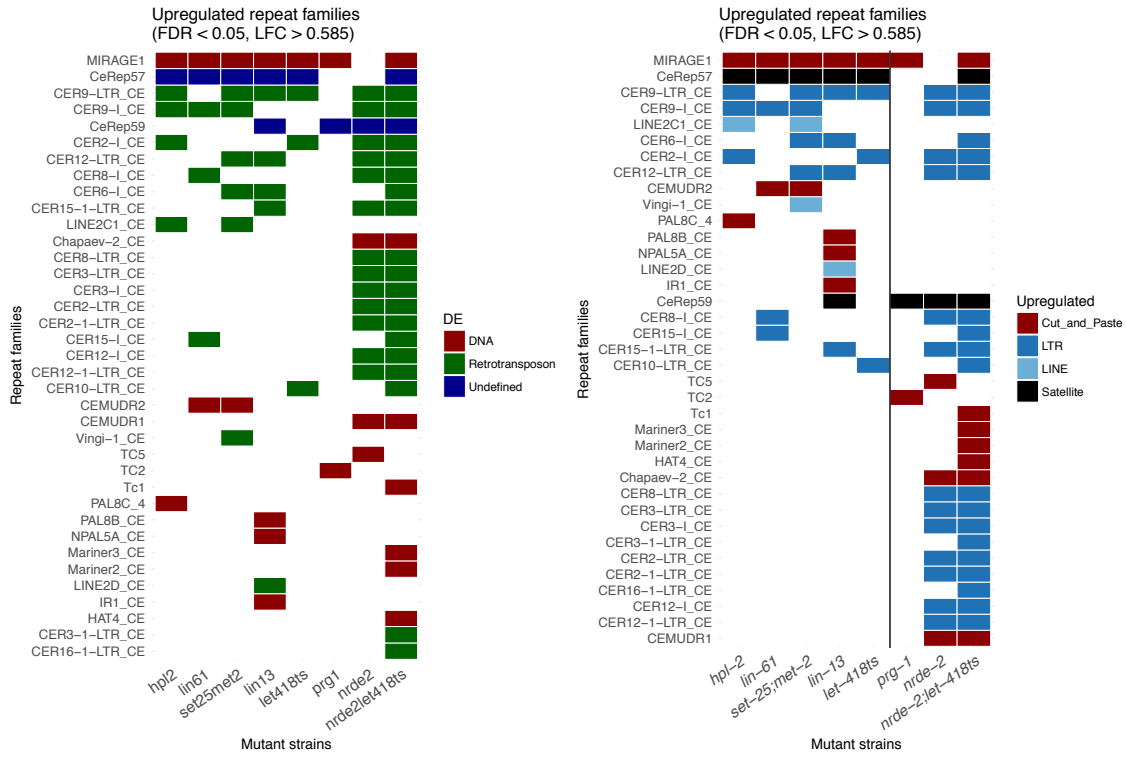


Figure 13 Summary of repetitive elements being de-repressed in heterochromatin mutants and small RNA pathway mutants. The left plot divides the repeats into DNA transposons, retrotransposons and undefined, while the right plot shows sub-division into functional classes.

2.5 Only a small fraction of repeats is de-silenced

As noted before only a very small fraction (<1%) of all repetitive elements is de-silenced in heterochromatin mutants or in *met-2 set-25* mutant that lacks H3K9 methylation. There are two possible reasons for this – most of repeats are not expected to be expressed, meaning that heterochromatin factors binding is not necessary required for silencing, and that there is a redundancy in function of 5 chromatin factors. I found

evidence supporting both models, and most likely, the extremely low fraction of repeats being up-regulated in chromatin mutants is due to a combination of both.

The transcription of most elements annotated in DEfam2.0 may not be actively supressed - only a small fraction of them have a potential to produce mRNA at all. 67% of elements are annotated as non-autonomous DNA transposons, which are mobilised in trans by a transposase. This transposase is usually encoded by a different gene annotated as an autonomous repeat. Furthermore, many individual repeats included in autonomous families are just short fragments and remnants of original repeats, that has lost the potential for being expressed – e.g. no longer have a promoter and ORF. Therefore, in my analyses I focused on transposons and retrotransposons – this set includes 221 repeats that overlap a predicted transposase ORF, plus 1085 LTR retrotransposons. The set of repeats upregulated in heterochromatin factors mutants is very strongly enriched for genes coding a transposase: 83-fold enrichment (21 of 71), and strongly enriched for LTR retrotransposons 10-fold (13 of 71). Almost half of all upregulated repeats encodes autonomous transposase. I conclude that the essential role of five heterochromatin factors is supressing repeats that have an autonomous function. The abundant binding of HC factors to repeats does not seem to regulate their expression in general. However, I speculate it might be important for other biological processes, for example preventing the binding of transposase, and thus preventing these elements to be activated in trans. It can also have a role in preventing homologous recombination at these loci, decreasing the chance of homologues recombination happening between repetitive elements, as has been reported before (Liu *et al.* 2011b; Sha *et al.* 2010). In conclusion, the heterochromatin factors are guarding genome integrity by (1) supressing expression from autonomous elements, (2) preventing activation in trans by autonomous elements, and (3) suppressing homologues recombination on repeats.

2.6 De-repressed repeats are weakly marked by heterochromatin factors

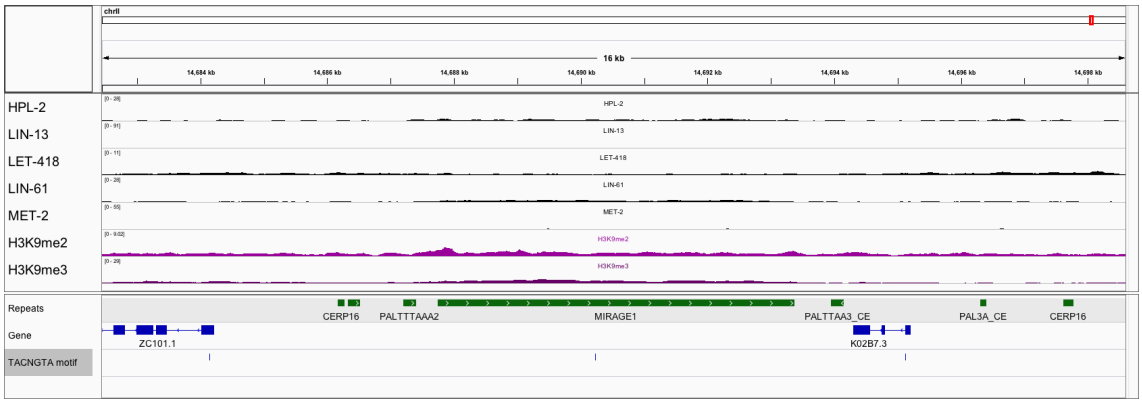


Figure 14 Very weak marking on MIRAGE1 element. This element is up-regulated in all heterochromatin mutant backgrounds. For comparative reasons all plots were scaled to **Figure 15** levels.

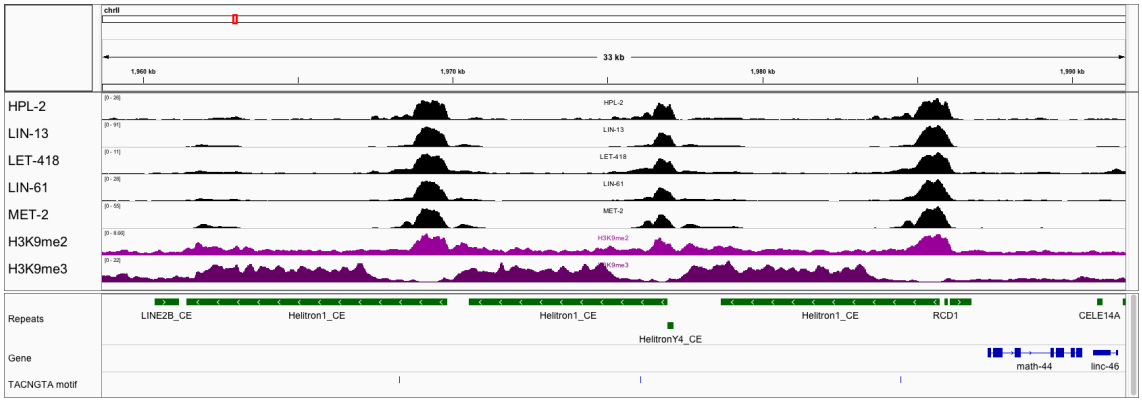


Figure 15 Promoter-like marking on Helitron1. Full length elements clustered on chromosome II are up-regulated in *lin-13* and *met-2 set-25* double mutants.

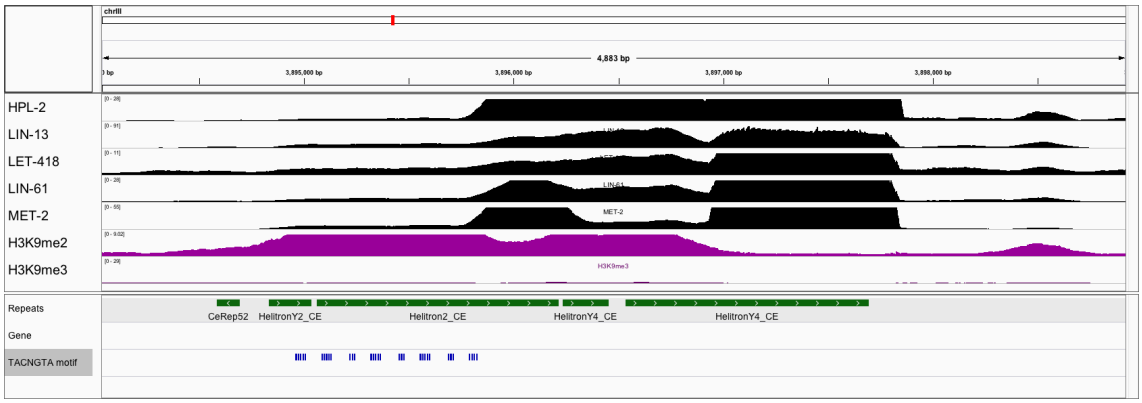


Figure 16 Very strong marking of all heterochromatin factors and H3K9me2 on Helitron2 and Helitron4. For comparative reasons all plots were scaled to **Figure 15** levels.

Transposase encoding repeats, that were de-repressed in chromatin mutants are in general weakly marked in comparison to repeats that stayed repressed. For example, MIRAGE1 is in general weakly bound by H3K9me2/3 and heterochromatin proteins (**Figure 14**). In contrast, Helitron1 element seems to have point marking of five factors and H3K9me2 upstream of its annotated TSS, which resembles a transcription factor binding locus – this configuration could suggest that it can be activated and may be regulated (**Figure 15**). Finally, Helitron2, which was not found overexpressed in any mutant background, has a strong marking of all five factors and strong marking of H3K9me2, but not H3K9me3, and H3K9me2 marks along the whole repeat body (**Figure 16**). The data suggest that heterochromatin factors have a collaborative, but partially redundant function in controlling the repeat expression, and repeats marked weakly have a higher chance to escape the suppression programme in a single mutant background. In such case the weak presence of other heterochromatin factors is not sufficient to maintain repressed state, and like in the case of MIRAGE1 I observe de-repression in all mutant backgrounds.

2.7 Heterochromatin factors mutant backgrounds show differential gene expression

In addition to repeats I have also analysed protein coding gene expression in heterochromatin mutants. I have found 267 to 513 genes being upregulated in individual mutant backgrounds. The most upregulated genes show three- to five-fold enrichment in expression in heterochromatin mutants compared to wild-type strains grown in the same temperature conditions. In addition to this there is a high overlap between genes upregulated in different mutant backgrounds – 404 out of 1155 genes are upregulated in more than two mutant strains. Also, heterochromatin factors are generally enriched on upregulated genes, but not downregulated genes, suggesting that the latter are indirect targets. Let-418 is an exception – it is enriched both on upregulated and downregulated

targets. Also, both H3K9me2 and H3K9me3 are enriched on upregulated genes in all mutant strains, suggesting that upregulated targets are indeed located in heterochromatin (**Figure 17**).

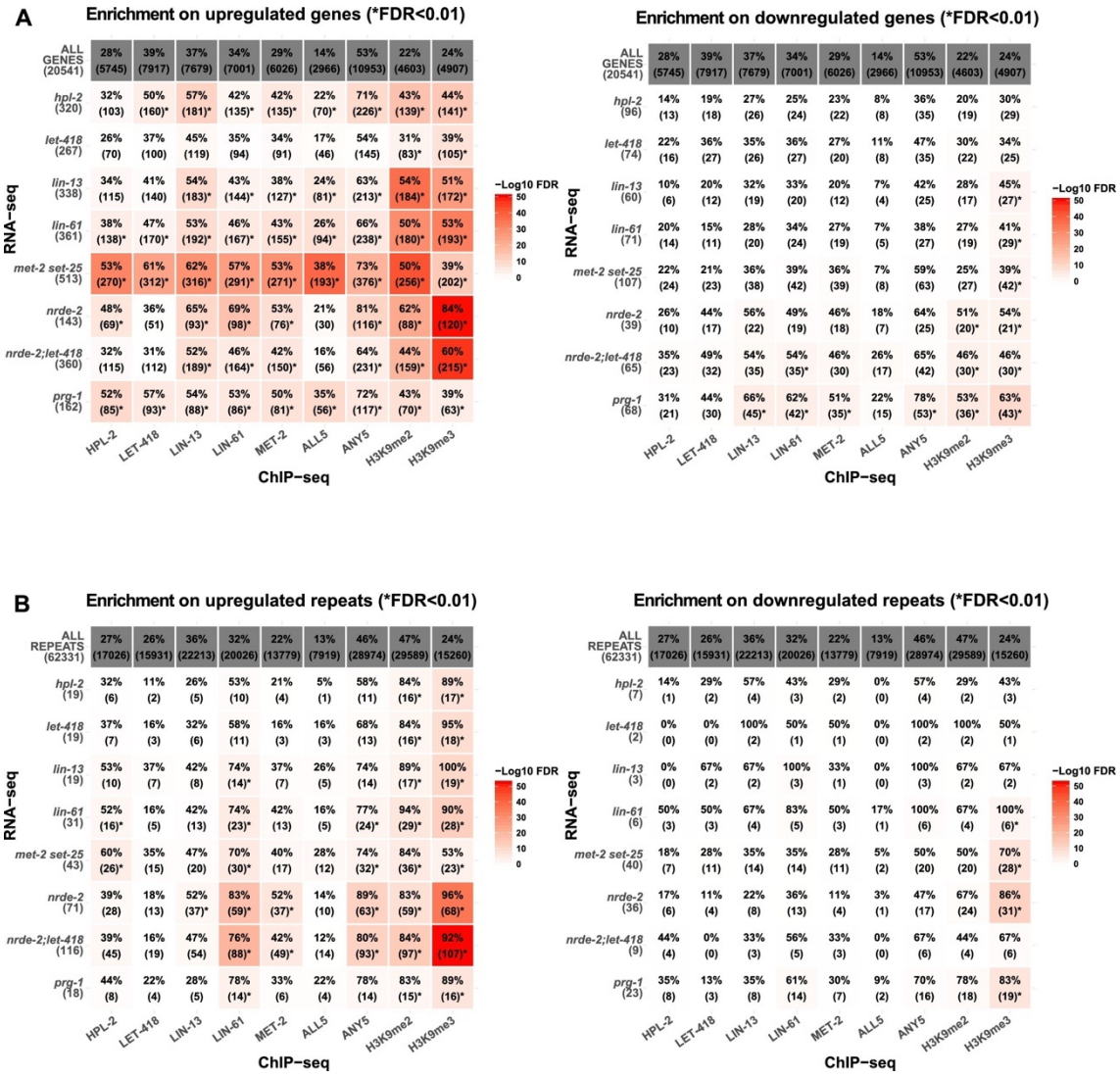


Figure 17 Heterochromatin factors and H3K9 methylation show enrichment on upregulated genes and repeats. (A) Percent overlap of heterochromatin factor peaks or >1.5 fold enrichment for H3K9me2 or H3K9me3 on upregulated (left) or downregulated (right) genes (−500 bp to gene end). Parentheses give number of genes with overlap, star indicates FDR < 0.01. (B) Enrichment of factors on upregulated repeats (left) and downregulated (right) repeats in each of the mutant strains.

Moreover, the enrichment is the strongest for genes upregulated in *met-2 set-25* double mutant, suggesting that H3K9me2 and H3K9me3 play direct role in gene expression regulation, or serve as binding platforms for proteins that repress expression, for example HPL-2 and LIN-61.

2.8 De-repression of MIRAGE1 elements partially causes the sterility phenotype

Since de-repression of repeats might provide stress to germline, I reasoned that it also may contribute the observed sterility phenotype. In further research I have focused on two autonomous repeats that were overexpressed, have a clear mechanism of actions described in the literature and encode an autonomous transposase. With my colleagues I have tested two hypotheses: (1) the MIRAGE1 elements cause double-strand breaks to DNA in donor and acceptor sites, and (2) the transposition of copy and paste HELITRON1 elements causes a single strand DNA overhangs at insertion sites that stalls replication fork causing replication stress. MIRAGE1 element is the most prominently de-repressed in all 5 investigated strains. There are 69 MIRAGE1 annotated in Dfam 2.0, of which only 6 are not truncated. The full-length elements contain two open reading frames, which encode ribonuclease H (RNase H)-like enzyme containing the active catalytic amino acid triad DDE/D. All these six, plus additional 6 truncated ones are upregulated in heterochromatin mutants. Both ORFs show similar levels of expression.

Using RNA-FISH my colleagues determined that wild-type adults had very low levels of MIRAGE1 RNA signal both in germ line and soma, which confirms my RNA-seq analyses. However, in *hpl-2, let-418*, and *lin-13* mutants we observed germ line localized MIRAGE1 RNA and almost no somatic RNA. This result confirms that HPL-2, LET-418, and LIN-13 repress MIRAGE1, and shows that signal detected in RNA-seq experiment originates from the germline.

Further, we wanted to determine if the overexpression of MIRAGE1 is indeed directly contributing to the sterility phenotype. My colleagues used RNAi to simultaneously target both ORFs of 16 different MIRAGE1 elements, including all untranslated occurrences in the genome, and most of truncated de-repressed ones (8/8 for *let-418*, 7/8

for *lin-13*, and 10/11 for *hpl-2*). RNAi was applied to *hpl-2*, *lin-13*, and *let-418* mutants, grown in nearly restrictive (causing sterility) temperature at 25°C, and led to small, significant increase of fertility in all cases. In addition, the improvement in somatic growth was observed. This result clearly illustrates that de-repression of MIRAGE1 contributes to the sterility phenotype. Taken all above observation together I conclude that heterochromatin factors maintain normal germline function and growth by repressing transposable elements that encode transposases.

2.9 The piRNA pathway shows similarity in repeat regulation and functional connections to heterochromatin factors

The most well studied and described mechanism of preventing transposon activity in the germline is the piRNA pathway (Weick *et al.* 2014). In *C. elegans,* the major player in this pathway is the Piwi Argonaut protein PRG-1, which facilitates silencing via both transcriptional, through engagement of the nuclear RNAi pathway, and post-transcriptional mechanisms, which are still poorly understood. Similarly to HC factor mutants, *prg-1* mutants show fertility defects, a low brood size comparing to wild-type, and mortal germline phenotype – all of these are more severe at higher temperatures (Batista *et al.* 2008; Das *et al.* 2008). In addition, recent reports suggest HPL-2 and H3K9 methyltransferase SET-25 are needed for piRNA pathway function in conjunction with the nuclear RNAi pathway (Ashe *et al.* 2012). All taken together, this suggests that HC factors silence repetitive elements together with proteins involved in piRNA pathway.

I started my investigation of connections between heterochromatin factors and PRG-1 by analysing RNA-seq expression profiling of *prg-1* mutant adults and focused on RNA produced by repetitive elements. I found only 18 elements to be de-repressed in *prg-1*

mutants. 14 out of 18 elements are also upregulated in at least one of the HC mutants, and this set includes MIRAGE1.

RNA FISH experiments performed by my colleagues confirmed the increase in MIRAGE1 RNA in *prg-1* mutant germlines – similarly to HC mutants (McMurchy *et al.* 2017). Next we checked if the germline apoptosis is also increased in *prg-1* background, and indeed we have observed a significant increase of germ cell death using the CED-1::GFP reporter strain. In conclusion, the phenotype, global expression profiling, RNA fluorescent *in situ* hybridization and apoptosis assay show similar results, suggesting that transposons are de-silenced in both *prg-1* and HC mutants. This means heterochromatin factors and piwi pathway (piRNA) have common silencing targets and might be acting together, guarding genomic integrity of the germline.

2.10 Heterochromatin factors act downstream of piRNA and subsequent 22G RNA synthesis

After establishing a link between HC factors and piRNA pathway, I wanted to ask if HC factors work downstream or upstream of piRNAs. Previous studies profiling small RNAs in *hpl-2* mutant background using piRNA sensor claimed that the abundance of both endogenous piRNA and ones targeting the sensor is not changed in comparison to wild-type (N2) strain (Ashe *et al.* 2012). This suggests that HPL-2 works downstream of piRNA generation. To investigated it further I have analysed 21U piRNAs and 22G siRNAs in HC mutant strains and in wild-type adults using following small RNA datasets from (Ashe *et al.* 2012): *prg-1* (GSM708661), WT matching *prg-1* (GSM708660), *hpl-2* (GSM950181), WT matching *hpl-2* (GSM950180), *nrde-2* (GSM950179), WT matching *nrde-2* (GSM950178).

A

	piRNAs		22G RNAs near piRNA target sites	
	piRNA number	repeats hit by piRNA	all 22G	Repeat associated 22G
WT	2853	4673	10237	3847
prg-1	36	53	5932	2074
WT	3561	7852	7388	2522
hpl-2	4282	10279	8011	2389

B

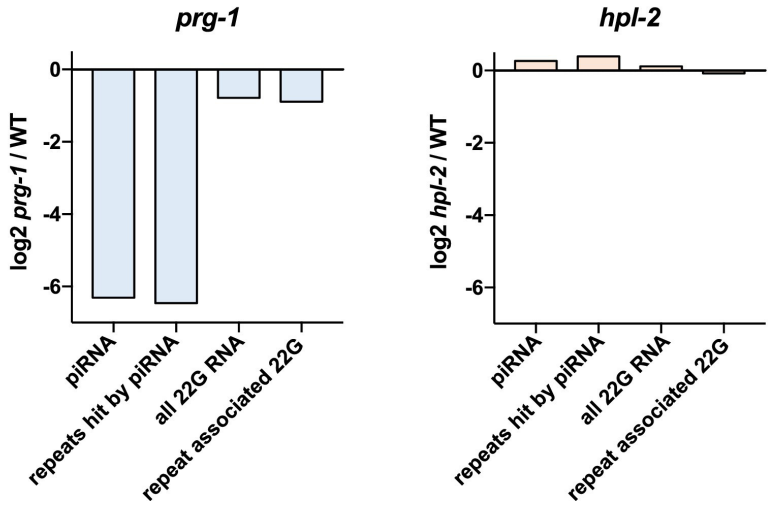


Figure 18 Summary of 21U piRNA and 22G siRNA abundance analyses. The table shows the number of unique piRNAs and 22G RNAs found in *prg-1* and *hpl-2* mutant background in comparison to the matching WT controls, and the number of repetitive elements that are predicted to be the targets of siRNA and piRNA. Below the same data is shown as the log2 fold change ratio between *prg-1* mutant and wild type.

I used *prg-1* mutant strain as the positive control for my pipeline for assessing the abundance of small RNAs, since PRG-1 loss-of-function impairs the production of piRNA, and in consequence the 22G siRNAs which are secondary to piRNAs. However, previous studies did not test a specific loss of piRNAs or 22G on repeats, so I also wanted to ask if repeats associated with piRNAs and 22G RNAs are also affected by *prg-1* mutant.

In agreement with previous reports (Lee *et al.* 2012), I detected almost complete loss of 21U piRNAs (79-fold depletion, 36 reads in *prg-1* in comparison to 2853 reads in WT). Also, the number of repeats targeted by piRNA dropped dramatically leaving only 53 out of 4673 repeats targeted in WT under control of piRNA pathway, representing 88-

fold depletion. Furthermore, the number of 22G RNAs mapping near predicted piRNA target sites also decreased from 10273 to 5932 in *prg-1* mutants, and the number of repeats targeted by 22G RNAs dropped from 3847 to 2074, representing 1.7-fold and 1.9-fold depletion respectively. It should be noted I did not expect a huge reduction of piRNA to directly translate to a similar reduction in 22G RNAs. piRNA pathway is only one way of generating 22G RNAs, and even the 22Gs coming from piRNA-dependant precursors can be maintained in *C. elegans* with the deficient piRNA pathway. Also, mapping between 21U piRNAs and 22G siRNAs is not trivial, since the exact process and abundance of this pathway, i.e. what percent of genome is scanned by piRNAs, remain unclear (Lee *et al.* 2012) (see Methods, chapters 6.1.12 to 6.1.15 for detailed implementation of the pipeline). These might provide a good explanation why I see so few repeats upregulated in *prg-1* mutant background despite the almost complete loss of piRNAs. 22Gs are main effectors of piRNA pathway, doing actual silencing both in cytoplasmic and nuclear pathway. Their population is depleted, but not as much as piRNAs, allowing only few repeats to escape their silencing programme. It makes it even more interesting that MIRAGE1 is one of these repeats.

Similar to the results cited in the beginning of this section, we found that the production of piRNAs in *hpl-2* mutants is not affected. The mutants show a small, possibly not significant increase in piRNA abundance, whereas the abundance of 22G siRNAs is not altered in *hpl-2* background. Hence, the interaction between piRNA pathway and HPL-2 appears to be downstream of piRNA and subsequent 22G RNA synthesis.

In conclusion, HPL-2, LIN-61, LET-418, SET-25, and MET-2 are important for the function of piRNA pathway. The collaboration between HC factors and piRNAs introduces a robust mechanism for supressing transposons, and possibly another layer of redundancy to the silencing system. Nevertheless, the abundant HC proteins binding

and de-silencing of more repeats in HC mutants in comparison to *prg-1* mutants indicate that HC factors facilitate repression that is outside of scope of piRNAs. This might include completely silenced repeats, as piRNA/siRNA pathway usually requires some low-level expression from its targets.

2.11 Nuclear RNAi pathway and heterochromatin factors are partially redundant

It is widely reported that the nuclear RNAi pathway in *C. elegans* facilitates transcriptional repression though directing H3K9me3 methylation to target genes (Alló & Kornblihtt 2010; Burkhart *et al.* 2011; Guang *et al.* 2010; Mao *et al.* 2015). The key player in this pathway is NRDE-2, and pathway is called nrde pathway, which stands for nuclear RNAi defective. Nuclear RNAi pathway acts downstream of piRNA generation and helps repressing piRNA targets (Ashe *et al.* 2012). To further investigate the interplay between HC factors, piRNAs and nrde pathway I undertook the expression profiling of repeats and genes in *nrde-2(gg91)* mutant using RNA-seq. *nrde-2(gg91)* is a putative null mutant, implying the complete loss of nrde pathway functionality. My analyses revealed that many more repetitive elements were upregulated in *nrde-2* background than in any of HC mutants or in *prg-1* background. They are also generally different elements than in HC mutants – out of 71 elements upregulated in *nrde-2* only 7 are also de-silenced in any of heterochromatin mutants. Outstandingly, none of MIRAGE1 elements upregulated in other tested backgrounds show de-repression in *nrde-2,* and DNA transposon targets are not enriched (**Figure 12B**). Contrary, the nrde-2 targets are enriched for retrotransposons, which represent 45 out of 71 targets. I conclude, that despite the dependence of piRNA pathway on nrde pathway to suppress its targets, the targets upregulated in these pathways deficient mutants differ. Another interesting observation is that the set of repeats de-repressed in *nrde-2* mutant and *set-25 met-2* double mutants is largely disjoint. *met-2 set-25* double mutant background

lacks detectable levels of H3K9me – this suggest that H3K9me might not be required for nrde pathway mediated repeats repression, or that lysine 9 methylation plays only supportive, rather than direct role in suppression of repetitive elements.

Similarly to heterochromatin mutants and *prg-1* mutant, *nrde-2* also shows temperature sensitive phenotype – the fertility is decreased at higher temperatures (Guang *et al.* 2010). However, considering previous results that different sets of repeats are de-repressed in HC and *nrde-2* mutants, it was not clear if there is a functional overlap in maintaining healthy germline and promoting fertility. To asses this I have analysed genes and repeats expression profiles in three newly constructed double mutants of *nrde-2* and three heterochromatin factors: *hpl-2*, *lin-61* and *let-418*. Since *nrde-2* mutant showed higher brood size reduction than any of HC mutants, the double mutants effects were assessed in comparison to it. We observed that there is no further reduction of fertility in *nrde-2; hpl-2*, and a weak, non-significant reduction for *nrde-2; lin-61*double mutants. However, *nrde-2; let-418* double mutant showed a significant reduction in the brood size in comparison to the *nrde-2* strain. This indicates that LET-418, but not two other tested heterochromatin factors, plays a partially redundant role in maintaining the fertility in *C. elegans*. Furthermore, my colleagues observed an increased embryo lethality in all 3 double mutants comparing to single ones, which supports classical additive phenotype for embryo lethality.

In order to understand if functional redundancy between LET-418 and NRDE-2 is connected to repeat suppression I analysed repeats profile in *nrde-2; let-418* adults using RNA-seq data. I found more repeats being de-silenced in double mutants than in any of the single mutants. 27 out of 46 elements were upregulated specifically in the *nrde-2; let-418* double mutant are retrotransposons. These data suggest that NRDE-2

and LET-418 have higher affinity to supress retrotransposons, while other HC factors and PRG-1 are more involved in supressing DNA transposons.

2.12 Repeat de-repression in *nrde-2* and *let-418* mutants happens in the germ line

Next question I wanted to tackle is if repeats miss-regulation happens only in germline or also in soma. I started this investigation by re-analysing previously published data - H3K9me3 ChIP-seq datasets in four different *nrde* mutants: *nrde-1*, *nrde-2*, *nrde-3*, and *nrde-4* (Buckley *et al.* 2012; Gu *et al.* 2009; Ni *et al.* 2014). I wanted to know if there was a difference in H3K9me3 marking on the genes and repeats upregulated in *nrde-2* mutants versus genes and repeats specifically upregulated in heterochromatin mutants. The nrde pathway represses its targets by marking the silenced loci with H3K9me3. I hoped to elucidate if NRDE-2 targets, in particular the ones common with LET-418 targets, show different H3K9me3 deposition and if this difference can be attributed to germ line vs. soma differences.

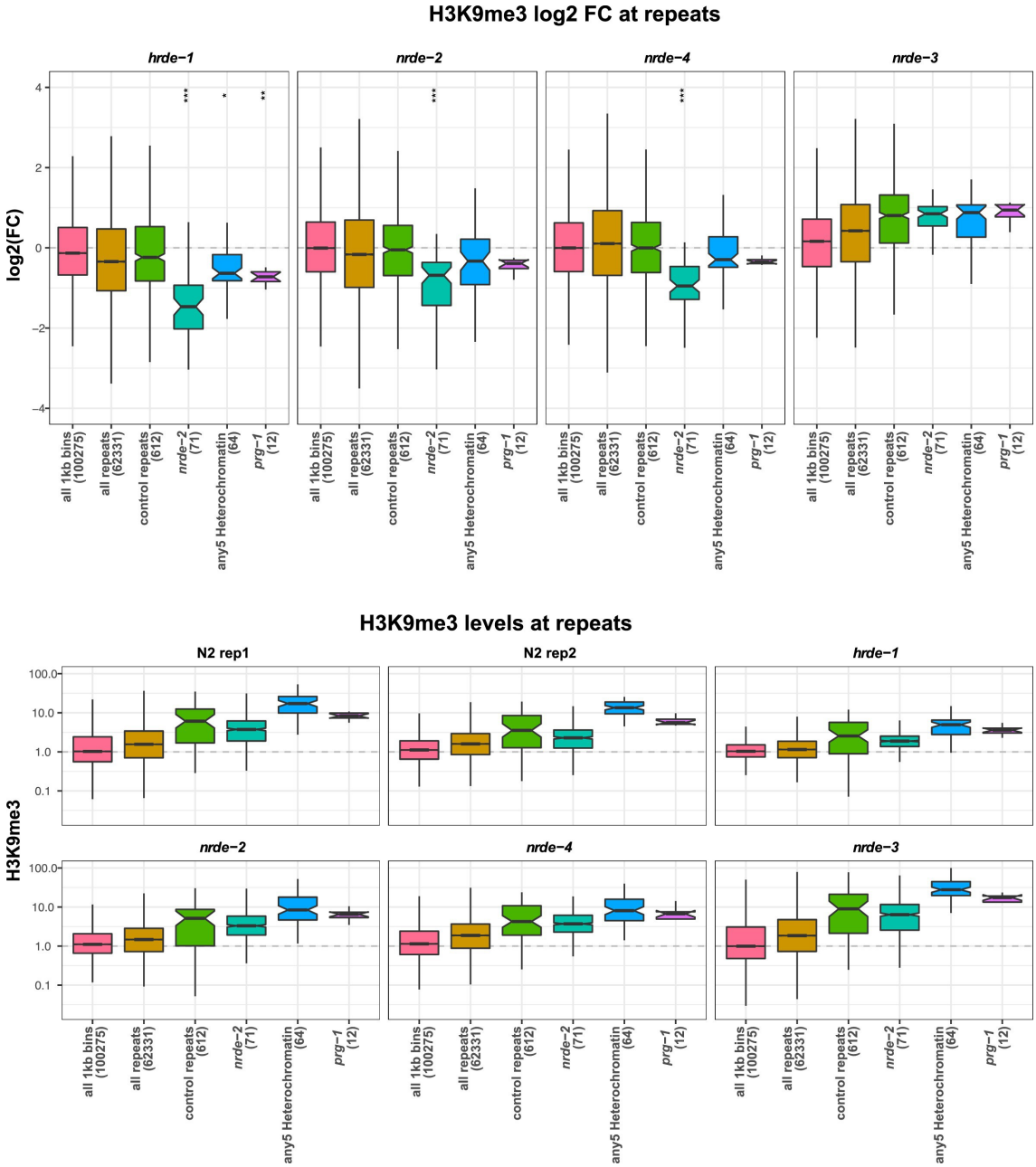


Figure 19 Boxplots showing gain or loss of H3K9me3 marking at indicated regions in *hrde-1*, *nrde-2*, *nrde-4*, and *nrde-3* background. The upper panel shows the log2FC of the average signal in indicated mutant relative to the wild-type signal. I defined a set of "control repeats" to have > 1.5 LFC enrichment relative to genome average, and no binding of heterochromatin factors. "nrde-2", "any5 Heterochromatin" and "prg-1" denote repeats up-regulated in *nrde-2,* any of five HC mutants, but not in *nrde-2*, and *prg-1,* but not in *nrde-2* mutant backgrounds, respectively. Parentheses indicate the number of elements in sets. Bottom plots show H3K9me3 levels relative to genome average at indicated regions. A reduction of H3K9me3 at repeat sets of interest was assessed by comparing to all repeats using a single-sided Mann-Whitney U test: one star, p<0.1; two stars, p<0.05; three stars, p<0.001.

This study takes advantage of the following assumptions: (1) HRDE-1 argonaut works only in the germline, (2) NRDE-3 works only in the soma, (3) NRDE-2 and NRDE-4

act in the whole organism (Buckley *et al.* 2012; Guang *et al.* 2010; Sarachana *et al.* 2010). The elements (genes and repeats) upregulated in *nrde-2* mutants also show reduced levels of H3K9me3. This shows that indeed the nrde pathway targets are supressed through H3K9 methylation. I observed a depletion of H3K9me3 at the same targets in *hrde-1* and *nrde-4* mutants, but not in *nrde-3*. This clearly shows that repeat upregulation due to deficient nrde pathway happens primarily in germ line, not in soma.

2.13 Repeats de-repressed specifically in heterochromatin factors mutants show small depletion in H3K9me3 in *hrde-1* and *nrde-4* mutants

Repeats upregulated in *hrde-1* and *nrde-4* mutant backgrounds showed reduced H3K9me3. This is similar to previously observed reduction of H3K9me3 repeats upregulated in heterochromatin mutants (*hpl-2*, *lin-13*, *let-418*, *lin-61*, *met-2 set-25*) and repeats specifically upregulated in *nrde-2* mutant background. The reduction in *hrde-1* and *nrde-4* was significant, but to a much lesser degree than in *nrde-2* mutant. I have not observed H3K9me3 reduction for repeats upregulated in in *nrde-3.*

There are two conclusions I draw from this result: (1) H3K9me3 levels at loci regulated by heterochromatin factors are controlled by some of germline nuclear RNAi pathways, and (2) small depletion of H3K9me3 at heterochromatin factor binding sites is not sufficient for de-silencing of repeats. This further supports a partial functional redundancy between heterochromatin factors and nrde pathway in repressing transposes.

2.14 Somatic repeats are de-silenced during aging process

In previous analyses I have elucidated the role of repeat element silencing in maintaining the fertility. As the analyses of H3K9me3 marking *hrde-1*, *nrde-2*, *nrde-3*, and *nrde-4* mutant backgrounds have shown that the silencing happens mostly in the germline, we can assume the de-repressed repeats are mostly of germline origin. This is not surprising, considering the affected mutants show the sterility phenotype. In conclusion, while profiling the repeats expression in adult *C. elegance,* vast majority of repeat expression profile will be contributed by germline.

However, heterochromatin mutants have also shown strong somatic growth defects, so it was of interest to profile the expression of somatic repeats. In order to assess the repeats expression profile in the soma I have profiled the expression of repeats in *glp-1* mutant background. This strain does not develop the germline, so all RNA-seq signal is of somatic origin.

2.14.1 Somatic repeat expression profiling in *glp-1* background

There is a good evidence of general de-repression and aberrant expression in aging (Tsurumi & Li 2012; Villeponteau 1997) and acute radiation poisoning (Sulli *et al.* 2012). Also, aging cells tend to accumulate DNA damage and have a weaker DNA repair machinery (Beerman *et al.* 2014). I aimed to investigate if repeats are also de-repressed in aging. Since a germline in *C. elegans* is essentially an immortal cell lineage with special mechanism for rejuvenation and maintaining a sturdy expression state, and as I established before the repeat expression profile in germline and soma significantly differs, I used the *glp-1* mutant to ensure I did study only somatic expression changes.

In this background I have analysed the expression profiles in aging time course, starting with the collection done as soon as worms reached maturity, named YA for young

adult, two days after YA collection (D03), six days after (D07), nine days after (D010) and thirteen days after (D14). YA is a time zero (T0) point in our analyses, all differential expression data are relative to this point (Janes *et al.* 2018). It should be noted that *C. elegans* lifespan is quite variable – wild-type (N2) hermaphrodites under apparently identical conditions (monoxenic plate culture, 20 C) can live from 11.4 to 19.9 days (50% survival or mean life span) (Johnson & Simpson 1985).

In this analysis I wanted to be extra sure that the aberrant expression of tested elements in aging is specific to repeats, rather than driven by the genes or gene outruns. Similarly, to previous analyses, I first analysed differential expression of genes using lax criteria (p-value <0.05, no fold change cut-off) and have filtered out repeats that overlap these gene bodies (both exons and introns). In addition to this, I have removed upstream outruns, based on most distant gene TSS annotated with short and long CAP RNA-seq data (Janes *et al.* 2018), and filtered out repeats that were closer than 500bp to annotated TSS of differentially expressed genes.

Condition	Filtered Repeats	Repeats	Genes	Repeat families
D03	76	134	1347	2
D07	283	472	2518	15
D10	421	659	2721	19
D14	578	1089	3492	27

Table 10 Differentially expressed genes and repeats in *glp-1* aging course.

As expected from previous reports, as the aging progresses, there was a steady increase of gene expression, starting with 1347 genes de-repressed in D03 (in comparison to YA), going up to 3492, which represents 2.6-fold enrichment. More strikingly, I found there is much stronger progressive de-repression of repeats, that are not in proximity to miss-regulated genes. In D03 there are only 76 de-repressed individual repeats, and this number goes up to 578 in D14. This represents 7.6-fold increase. Also, the number of de-repressed families increases during aging process – the 76 individual repeats

upregulated in D03 represent only 2 families. However, on D14 there are 27 families upregulated, which represent 13.5-fold increase (**Table 10**). In conclusion – both genes and repeats are de-repressed in aging, but the gene-independent repeats show a much more striking increase.

2.15 Repeats de-silenced in aging remain expressed throughout further aging course

After establishing that repeats are de-repressed in aging I wanted to understand if the repeats that got turned on at certain stage stay expressed throughout aging course. It is important, because stochastic turning on or off of the expression of repeats would suggest that previous observation just captured random aberrant expression. Contrary, if the same repeats stay expressed through aging course it would suggest they might have functional relevance and do contribute to aging phenotype, e.g. by contributing to DNA damage accumulation.

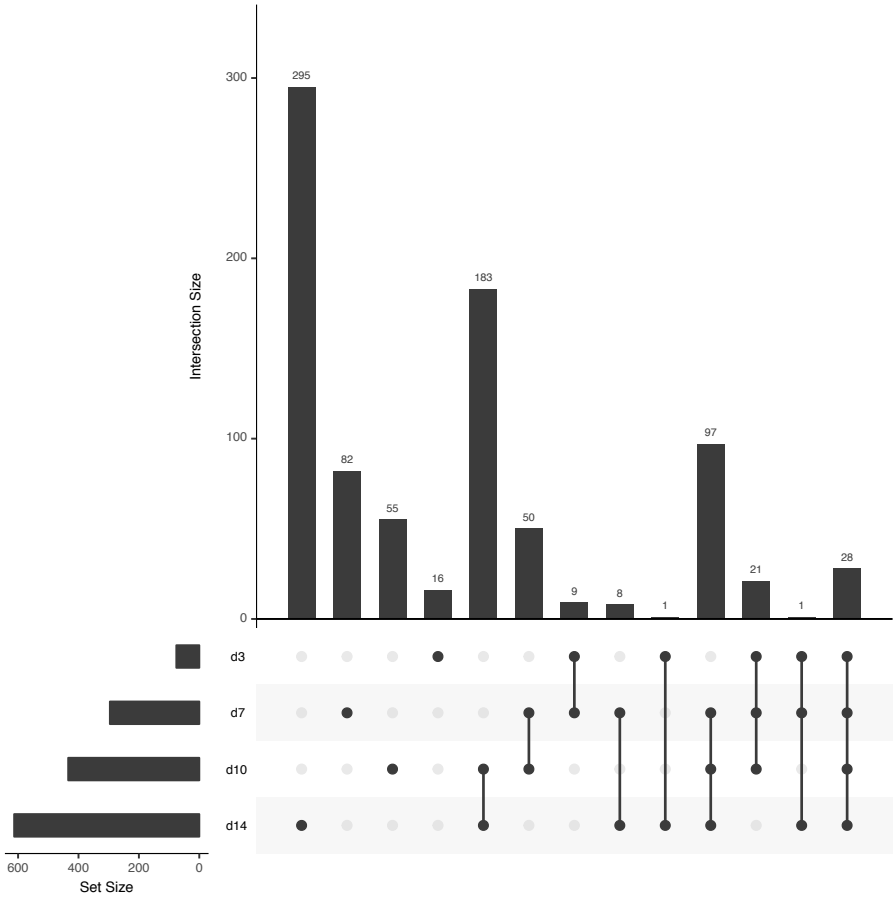


Figure 20 The UPset plot showing the overlap between significantly upregulated repeats in D03 to D14 time points.

Analyses of overlaps between the significantly upregulated repeats in aging course time points (**Table 10**) showed that 37% (28 out of 76) repeats upregulated in D03 were found upregulated in all later time points, and only 16 repeats were not found upregulated later (**Figure 20**). For D07 time point 33% (97) repeats and for D10 42% (183) repeats were found upregulated in all later time points. 295 repeats were de-repressed in the last time point – D14. This data is showing that a larger fraction of repeats de-repression is carried through ageing. However, I thought that fluctuations in repeat expression between aging time points might render some repeats expressed in relatively stable manner not passing the significance threshold (FDR <0.01).

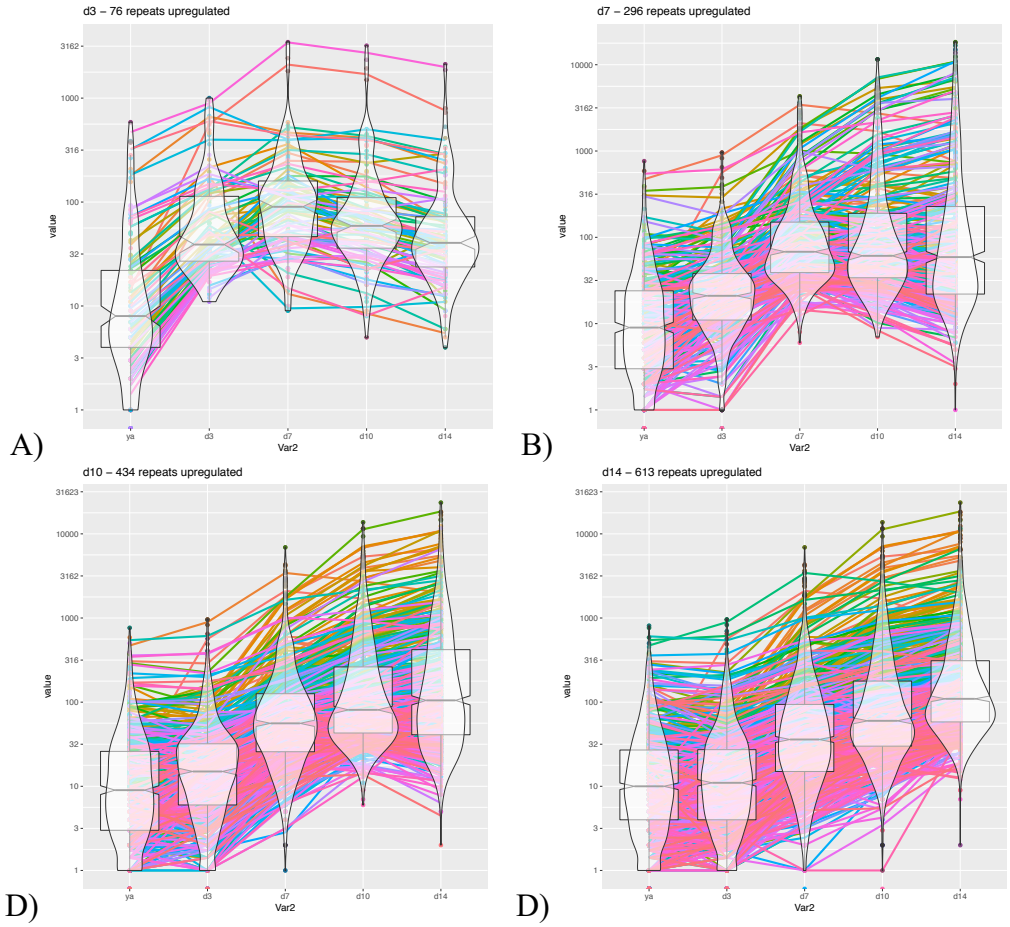


Figure 21 The panels show expression estimates of repeats throughout aging course. A) repeats significantly de-repressed three days after YA collection (D03) comparing to YA, B) seven days after (D07), C) ten days after (D010) and D) fourteen days after (D14). The lines are coloured randomly to allow easier tracking of individual genes.

For this reason, I decided to track individual repeats upregulated at a given time point through the whole time course – from YA to D14. Some of the repeats that are de-repressed in D03 revert to lower expression in later stages, but generally their expression in later stages is significantly higher than in YA (**Figure 21A**). Almost all repeats de-repressed in D07 and D10 remain highly expressed until day D14 (**Figure 21B** and **C**). Further, I tracked the expression history of repeats upregulated in day D14: on D03 the median expression is only slightly higher than in YA, but there is a significant and sturdy gain in median expression in D07 and D10 (**Figure 21D**).
I conclude, that there is sturdy and progressive de-silencing of repeats in *C. elegans* aging - once the repeat got de-silenced it maintains its expression, and repeats overexpressed in D14 time point show a sturdy growth in expression at previous time points. This suggests that repeats might be contributing to aging process.

2.16 TAC(B)GTA motif is enriched on repeats and HCF binding sites

Next, I asked if there is a particular motif or set of motifs connected to repetitive elements. Using MEME motif searching software I found that a set of motifs is enriched in the repetitive elements associated and all of chromatin factors. Particularly, I found the pseudo-palindrome TAC(B)GTA motif (**Figure 22** A and B) to be highly enriched with E-value ranging from 9.7e-440 (LIN-13) to 2.6e-141 (MET-2). It co-localizes with factor peaks on Helitron1 or decorates Helitron2 - a repeat highly enriched for all heterochromatin factors binding. Furthermore, the motif is present on many miss-regulated repeats, for example, all full length, autonomous MIRAGE1 (all misregulated - 7) and Helitron1 (all misregulated - 5) repeats. Remarkably, out of ~35K motif sites genome-wide 19249 and 12016 overlaps with LIN-13 broad and narrow peaks, respectively. Considering that peaks constitute very small fraction of genome, the enrichment over randomly distributed motif sites is outstanding. Beside the primary

TACBGTA motif, I have found a secondary motif RTASGCA, which is interspaced by 9bp of less conserved DNA. This might suggest that the proteins recognising TACBGTA often co-localise with other factor or factors recognising the RTACGCA motif. The secondary motif along the primary one is shown on **Figure 22**C.

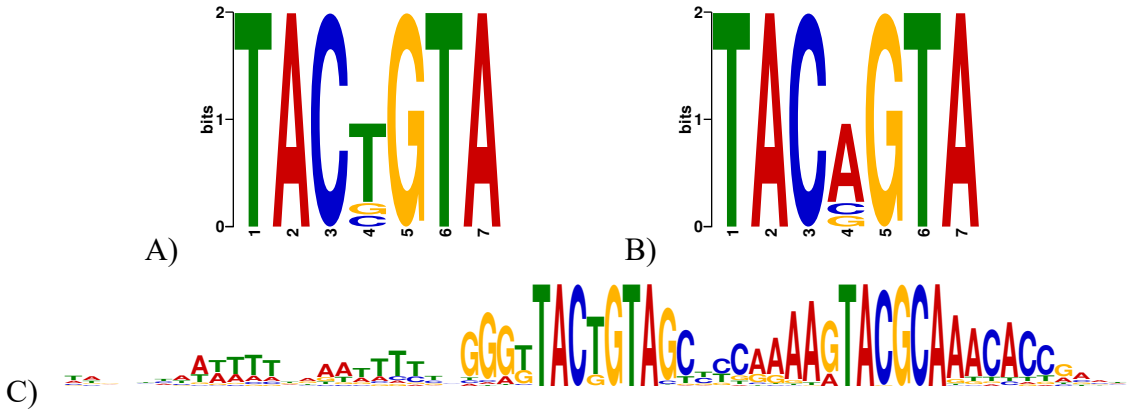


Figure 22 TAG(B)GTA motif logo in normal and reverse complement orientation (above, A and B) – note the motif is a pseudo-palindrome; secondary motif RTASGCA shown along with primary one (below, C)

The motif enrichment is potentially a very important observation that may lead to a better understanding of the heterochromatin factor binding mechanics. Despite the fact LIN-13 binding loci are strongly enriched for TAC(B)GTA, only around 35% of LIN-13 peaks have the motif. This suggests that TAC(B)GTA may not be directly involved in LIN-13 binding to DNA, and there might be another DNA binding protein directing the chromatin factors binding.

To further investigate this possibility, I found mammalian DNA binding proteins known to bind a similar motif: FOXI1 forkhead family transcription factor, and Osr1 and Osr2 - odd-skipped-related C2H2 zinc fingers proteins. FOXI1 has no orthologs in *C. elegans*, but OSR1 and OSR2 have 2 orthologs: ODD-1 and ODD-2.

There is very limited literature on *odd-1* and *odd-2* genes, but from the published data I have found that there is a phenotypic similarity to chromatin factors - ODD-1 is

involved in gut development during embryogenesis (similarly to LET-418) and affects larval viability. It is also expressed at the same development stages as five heterochromatin factors. This suggests that ODD-1/2 might be a new heterochromatin player, facilitating the binding of other heterochromatin factors. Further experiments are required to test this hypothesis.

2.17 Possible source and functions of TAC(B)GTA motif

The high copy number of TAC(B)GTA motif might be connected to evolutionary history of repeats transposition - TAC(B)GTA is excision footprint for Tc1/Mariner superfamily members: Sleeping Beauty and Frog Prince transposons. Sleeping Beauty (SB) leaves exactly TACAGTA/TACTGTA excision footprint, while Frog Prince (FP) shows imprecise excision, generally plus 5 base pairs from TAC(A/T)GTA footprint (Hucks 2008; Ni *et al.* 2008). Furthermore, this motif is a pat of miRNA seed sequence for three micoRNAs: miR-101, miR-199a, miR-144:

- miR-101 seed sequence: TACTGTA - targeting CFTR 3'UTR located microRNA responsive elements (MRE) (Megiorni *et al.* 2011)
- miR-144 seed sequence: ATACTGT (van Dongen *et al.* 2008)
- hsa-miR-101 hsa-miR-199a* hsa-miR-144 and seed sequence: TACTGTA (Corà *et al.* 2007)

Also, TAC(B)GTA is a type II restriction enzyme recognition sequence for SnaBI enzyme – isoschizomers XcaI, BstSNI, EciAI and Eco105I (Brown 1998). Interestingly, similar motif was found in bacterium operator. DNA operator in *Streptomyces lividans*, showing A-DNA-like topology, controls copper sensitive operon repressor (CsoR). CsoRSl binds to its operator site through a 2-fold axis of symmetry centred on a conserved 5'-TAC/GTA-3' inverted repeat. Circular dichroic signatures of the CsoRSl–DNA interaction suggest selectivity towards the A-DNA-like topology of the G-tracts at the operator site. Differential binding modes may exist in operator sites

having more than one 5'-TAC/GTA-3' inverted repeat (Tan *et al.* 2014). Considering far evolutionary distance between *S. lividans* and *C. elegans* these regulatory mechanisms might not be related. However, it is interesting to see similar sequences preforming regulatory roles in bacteria and metazoa.

2.18 Distribution of repeats in *C. elegans* genome

To better understand what biological processes might be influenced by aberrant repeat expression, I studied how repetitive elements are distributed in *C. elegans* genome. Dfam 2.0 annotates 62,331 repetitive elements, covering 16.26 million base pairs - 16.21% of *C. elegans* genome. As mentioned above in the context of repeat expression analyses, repeats are more densely located on chromosome arms. Chromosome arms are enriched for DNA transposases and uncharacterised/satellite repeats, while central regions show higher retrotransposon enrichment.

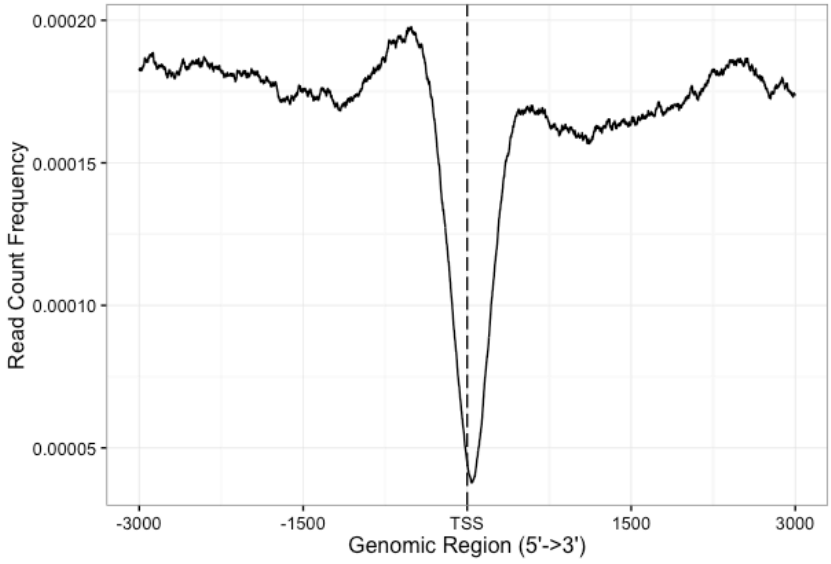


Figure 23 Repeats are depleted on TSS of coding genes. The graph is based on combined CAP RNA-seq/Wormbase annotation of TSS and covers the region +/- 3kb from TSS.

Repeats are depleted on TSS, showing 4-fold depletion in comparison to proximal (+/- 3kb) background, and only slightly depleted (<1.5-fold) in gene bodies (**Figure 23**). Interestingly, there is a slight enrichment of repeats around 500bp upstream of TSS.

This is consistent with a theory that many regulatory regions originate from domesticated repeats and still possess repeats characteristics (Ward *et al.* 2013).

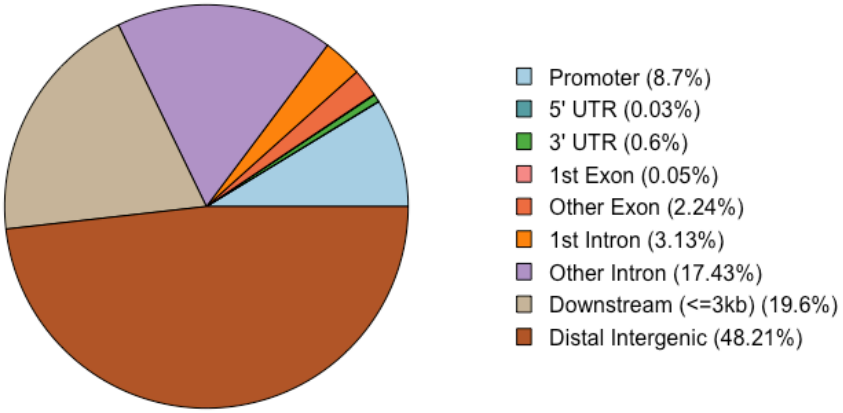


Figure 24 The pie chart showing distribution of repeats in *C. elegans* genome.

In terms of genomic location, almost half (48.21%) of repeats maps to distal intergenic regions. 20.47% map to introns, with 3.13% in the 1st intron and 17.34% in other introns. Further 19.6% map to downstream gene regions (<= 3kb downstream annotated translation termination site). Despite strong TSS depletion, 7.8% repeats map to promoter regions (defined here as up to 1kb upstream TSS), which further supports their role in creating regulatory regions. Repeats are depleted from exons, with only 0.05% mapping to 1st exon and 2.24% mapping to other exons of protein coding genes. Also, untranslated regions are particularly depleted, with only 0.03% mapping to 5'UTR and 0.6% mapping to 3'UTR (**Figure 24**).

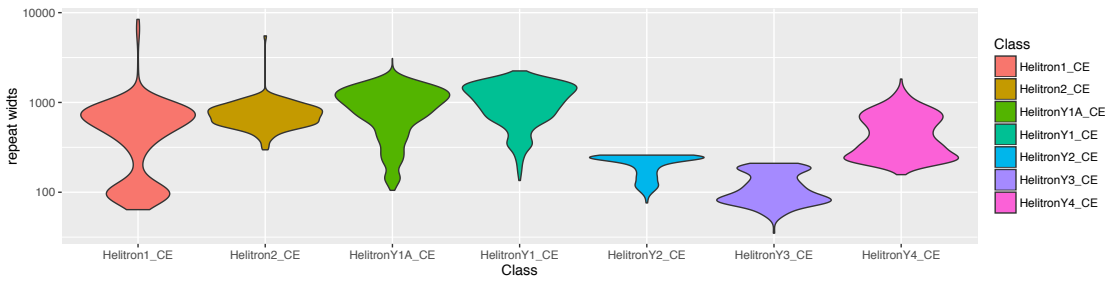


Figure 25 Violin plots showing length distribution of top 10 most abundant repeats.

Further investigation of 62,331 repetitive elements revealed that the vast majority of repeats annotated as DNA transposon and retrotransposon are short remnants of an original repeat. Only a very small fraction of repeats encodes full functional transposon, for example, for Helitron1 repeat there are only 5 out of 432 annotated repeats. Repeats are very diverse and within a single family show diverse lengths (**Figure 25**), enrichment in heterochromatin factors, as well as the potential to be de-silenced in heterochromatin mutants. **Figure 26**, **Figure 27** and **Figure 28** show examples of different classes of repeats: cut and paste DNA transposon - PALTTAA2, dispersed repetitive element - CEREP5, and SINE retrotransposon - LmeSINE1c.

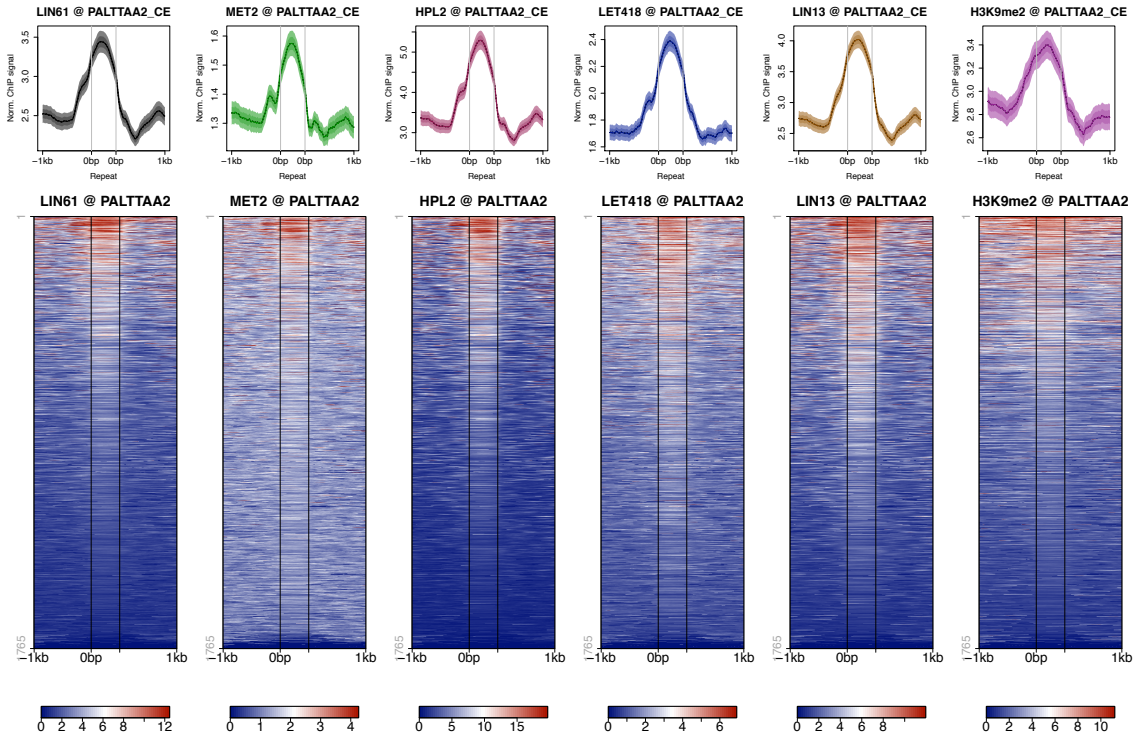


Figure 26 Chromatin marks on PALTTAA2 CE repeats.

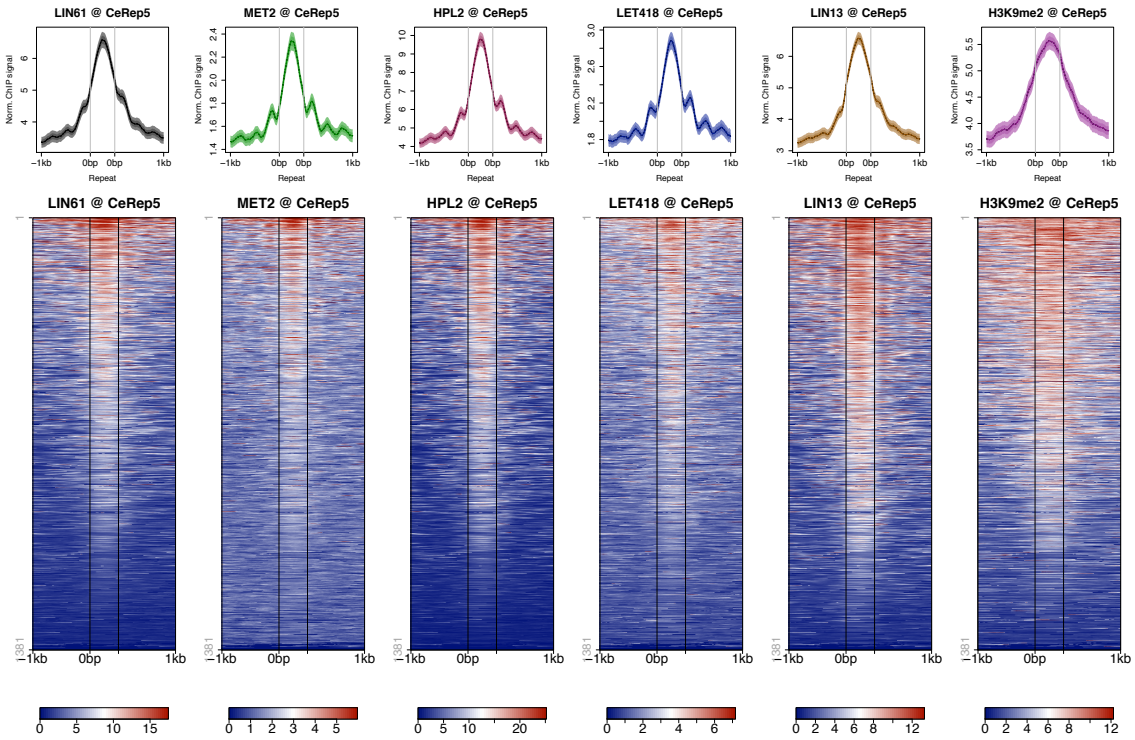


Figure 27 Chromatin marks on CEREP5 repeats.

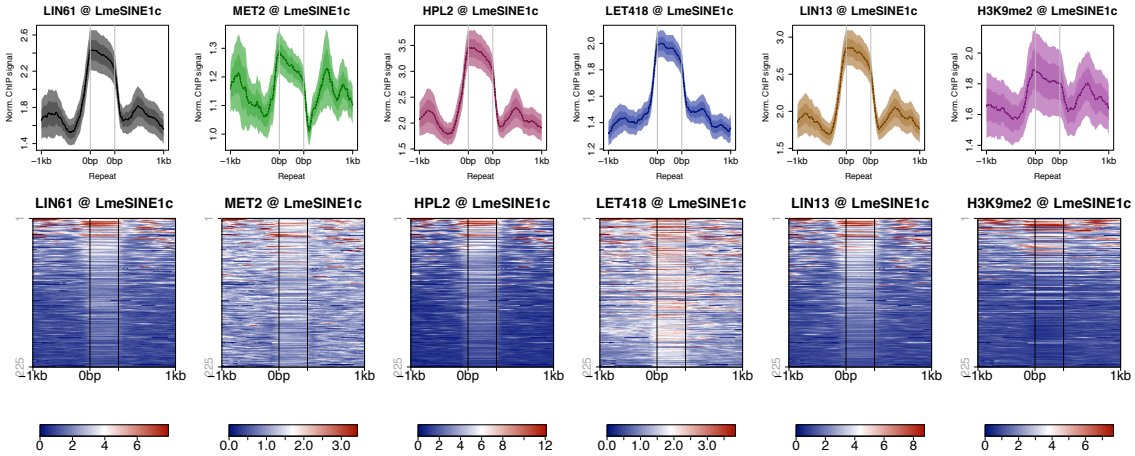


Figure 28 Chromatin marks on LmeSINE1c repeats.

Another, very important feature of *C. elegans* repeats is that they are distributed as single repeats, or in small domains interspaced by unique DNA. This is in contrast to *H. sapiens* or *D. melanogaster* models. This feature makes *C. elegans* a very good model to study repetitive elements for two reasons: 1) the genome assembly has quite accurate and complete sequences of most repeats and sequencing based experiments, such as RNA- and ChIP-seq, usually provide enough unique reads on repeat – unique DNA

junctions to pinpoint the binding or differential expression event to particular, single repeat loci, and 2) repeats in higher animals, and H3K9me3 associated with them, usually play a role in centrosome formation.

2.19 There is no defect in splicing in *hpl-2* mutant strain

Both repetitive elements and five analysed heterochromatin factor peaks are commonly located in introns of *C. elegans* genes. This observation prompted me to investigate if HC factors have any role in controlling splicing machinery in *C. elegans*. It should be noted that though *C. elegans* genome is very small and compacted comparing to mammalian genomes, intergenic distances constituting a similar fraction of the genome as introns (**Figure 24**). Consequently, there is much higher relative fraction of intronic repeats in *C. elegans* comparing to organisms with less compacted genomes. This observation points me towards an alternative hypothesis, that heterochromatin factors have no role in splicing, and are associated with intronic repeats, potentially suppressing abradant transcription.

EventType	N JC only	Sig JC only	N JC + readsOnTarget	Sig JC + readsOnTarget
Skipped exon (SE)	1404	9 (5:4)	1449	14 (11:3)
Mutually exclusive exon (MXE)	129	4 (2:2)	137	2 (0:2)
Alternative 5' splice site (A5SS)	302	7 (4:3)	304	7 (3:4)
Alternative 3' splice site (A3SS)	642	7 (3:4)	643	7 (3:4)
Retained Intron (RI)	332	13 (7:6)	333	14 (7:7)

Table 11 rMATS results *hpl-2* vs N2 differential splicing analysis shows very few events found on the global scale. Columns denote: *N JC only* - total number of events detected using Junction Counts only. *Sig JC only* - number of significant events detected using Junction Counts only. The numbers in the parentheses (n1:n2) indicate the number of significant events that have higher inclusion level for *hpl-2* (n1) or for WT (n2). *N JC + readsOnTarget* - total number of events detected using both Junction Counts and reads on target. *Sig JC readsOnTarget* - number of significant events detected using both Junction Counts and reads on target. The numbers in the parentheses (n1:n2) indicate the number of significant events that have higher inclusion level for *hpl-2* (n1) or for WT (n2).

I observed no global effect on splicing in heterochromatin mutants using two independent computational tools MISO and rMATS (**Table 11**) in WT vs. *hpl-2* mutant experiment. Differential splicing analyses produced very few significant events. I validated them all by browsing in IGV genome browser and found out they were very likely to be false positives (**Figure 29**). In addition to this, none of the events overlapped known repeat or chromatin factor binding sites. Furthermore, I have run the alternative software MISO. The results were similar – I have found 13 differential splicing events to be significant in both replicates (Bayesian factor > 150), and after validation by plotting Sashimi plots (**Figure 30**) (Katz *et al.* 2015) and browsing in IGV I discovered many to be false positives. None of the identified events overlapped heterochromatin factor binding sites. These data suggest there is no evidence for global splicing defects in *hpl-2* mutant background.

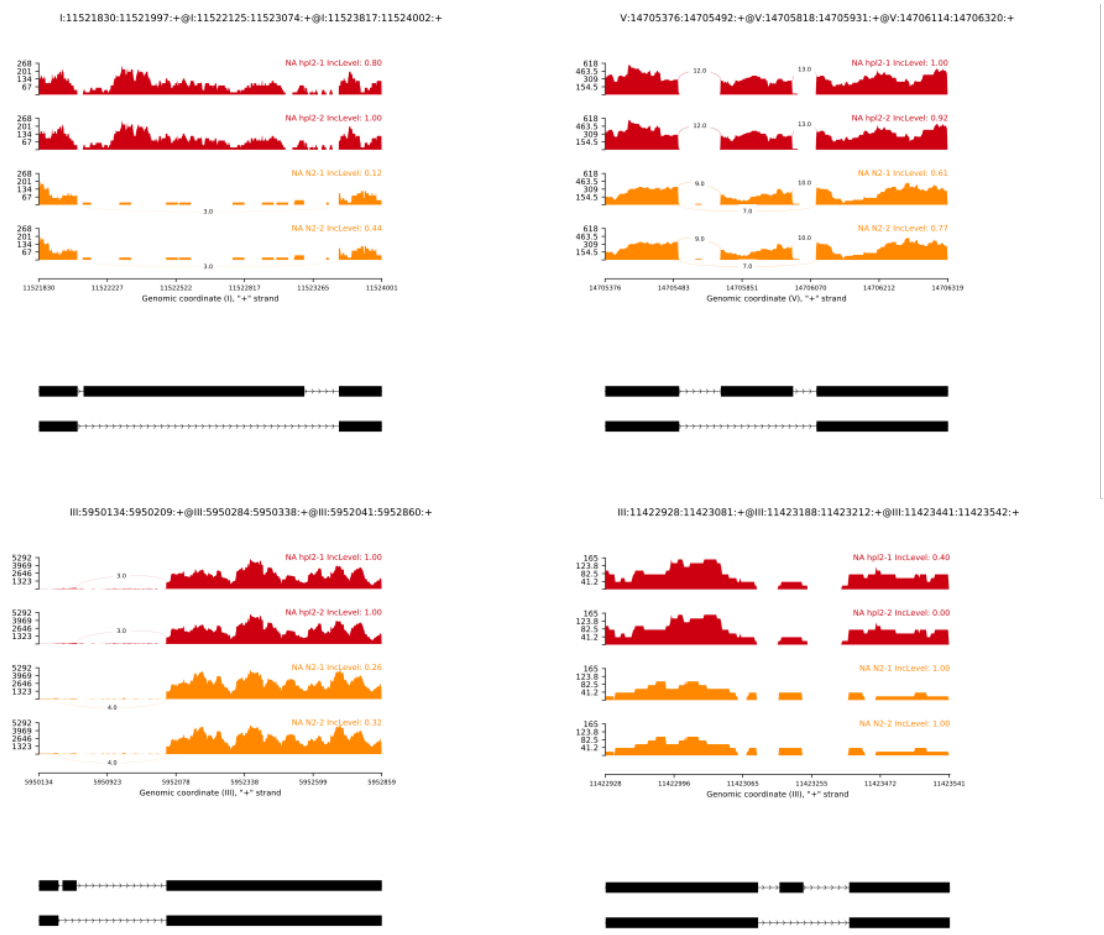


Figure 29 Splicing events found with rMAST are most likely false positives. The plot shows example loci of differential significant spicing events.

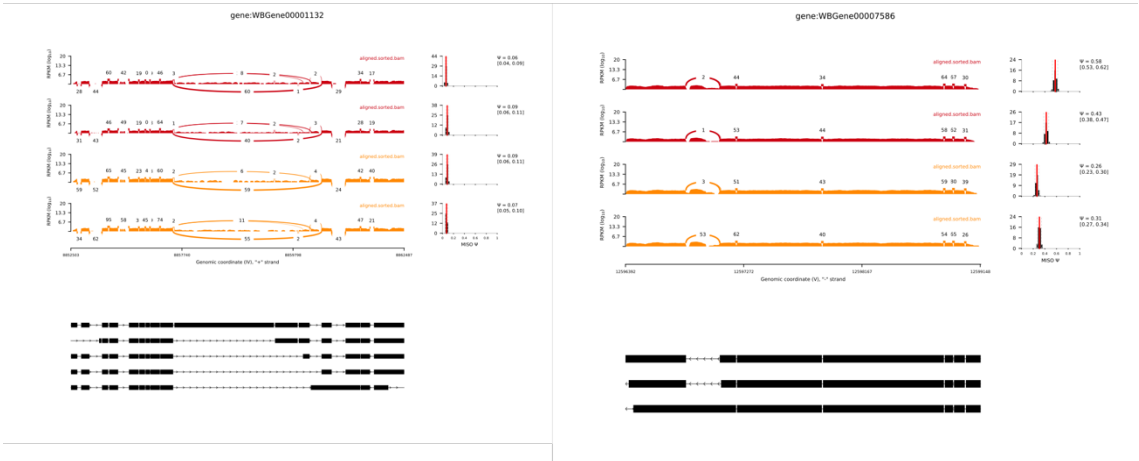


Figure 30 Splicing events found with MISO are most likely false positives. The plot shows example loci of differential significant spicing events.

3 PROMOTERS AND OPEN CHROMATIN

In Chen et al, 2013, we defined *C. elegans* promoters and enhancers for the first time by profiling transcription initiation and transcription elongation in nuclear RNA. We observed that promoters and enhancers share properties. They both bind RNA polymerase, initiate transcription, and overlap mapped transcription factor binding sites (TFBS). Other studies in *C. elegans* and other organisms showed that some TFBS had the unusual property of binding a large number of factors (called highly occupied targets, or HOT regions), but their function was unknown. In this chapter, I first describe analyses of HOT regions in *C. elegans* and humans, in which we showed that they are ubiquitously active CpG dense promoters. I then describe further analyses of *C. elegans* promoters.

HOT (high occupancy target) regions are an interesting class of elements found in *H. sapiens, D. melanogaster* and *C. elegans* genomes. In ChIP-seq assays, HOT regions are observed to show binding of most known TFs. The number of factors binding in these locations is much higher than expected from negative binomial distribution. Further, these regions generally lack classical, sequence-specific TF binding motifs for immunoprecipitated factors. These characteristics show they are not a classical enhancers (Gerstein *et al.* 2010; Roy *et al.* 2010). Also, it was previously reported in *D. melanogaster*, that HOT regions are not generally overlapping annotated enhancers.

However, when tested using transgenic reporter assay many HOT regions displayed enhancer activity (Kvon *et al.* 2012). In both *D. melanogaster* and *H. sapiens* HOT regions display open chromatin characteristics, such as nucleosome depletion and high nucleosome turnover, and in C. *elegans* they tend to be located upstream of ubiquitously expressed genes (Gerstein *et al.* 2010; Roy *et al.* 2010; Yip *et al.* 2012). These reports show that HOT regions have a solid hallmark of regulatory elements.

3.1 Transcription factors overlap extensively in *C. elegans* and *H. sapiens* in high occupancy of targets (HOT) sites

To define HOT regions in *C. elegans* and human, I used collections of transcription factor (TF) mapping data from modENCODE and ENCODE consortia (Birney *et al.* 2007; Boyle *et al.* 2014; Gerstein *et al.* 2010, 2012; Nègre *et al.* 2011; Niu *et al.* 2011; Roy *et al.* 2010). At the time, 90 *C. elegans* factors and 159 human factors had been mapped. I collected the published peak calls and determined the regions where factors overlap (**Figure 31**). Because the datasets were collected from multiple cell types and developmental stages, some overlaps may not actually happen in a single cell type. Nevertheless, the overall number of factors binding in given loci is a good proxy of the occupancy of targets.

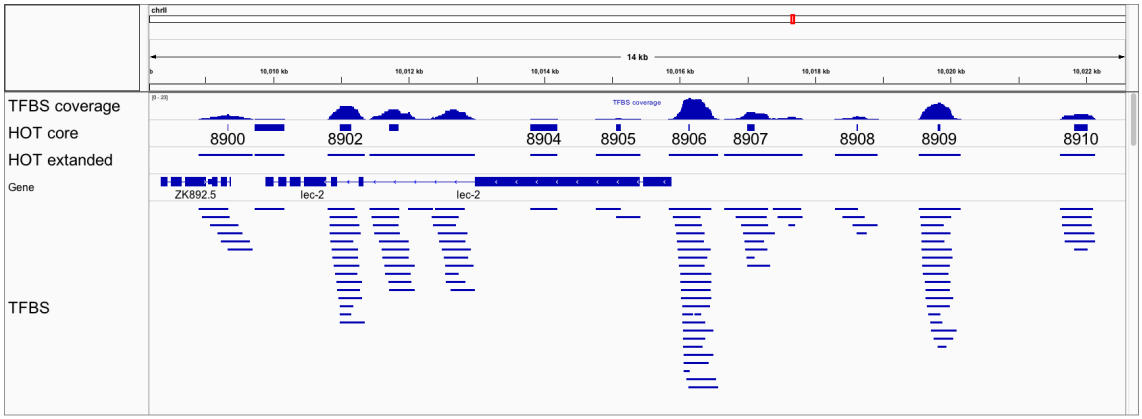


Figure 31 Genome browser screenshot showing the principal of HOT region assignment. Transcription factor binding sites are shown in the bottom. Top track shows summarised coverage of TFBS. The covered regions are assigned as HOT extended, while the summits are assigned as HOT core.

Each region was scored for the number of unique factors binding in a given location. I identified 35,062 overlapping regions in *C. elegans* bound by 1–87 factors and 737,151 overlapping regions in *H. sapiens* bound by 1–138 factors. There is a wide range in the number of TFs per site in both human and *C. elegan*s. Most sites bind only a single or just a few transcription factors, as expected of classical enhancers. However, in the top 1% of occupancy, 57/138 or 64/87 factors bound in humans or *C. elegans*, respectively. Below, I analyse the function, DNA sequence and chromatin characteristics that distinguish HOT regions from other regulatory regions.

3.2 HOT regions are promoters

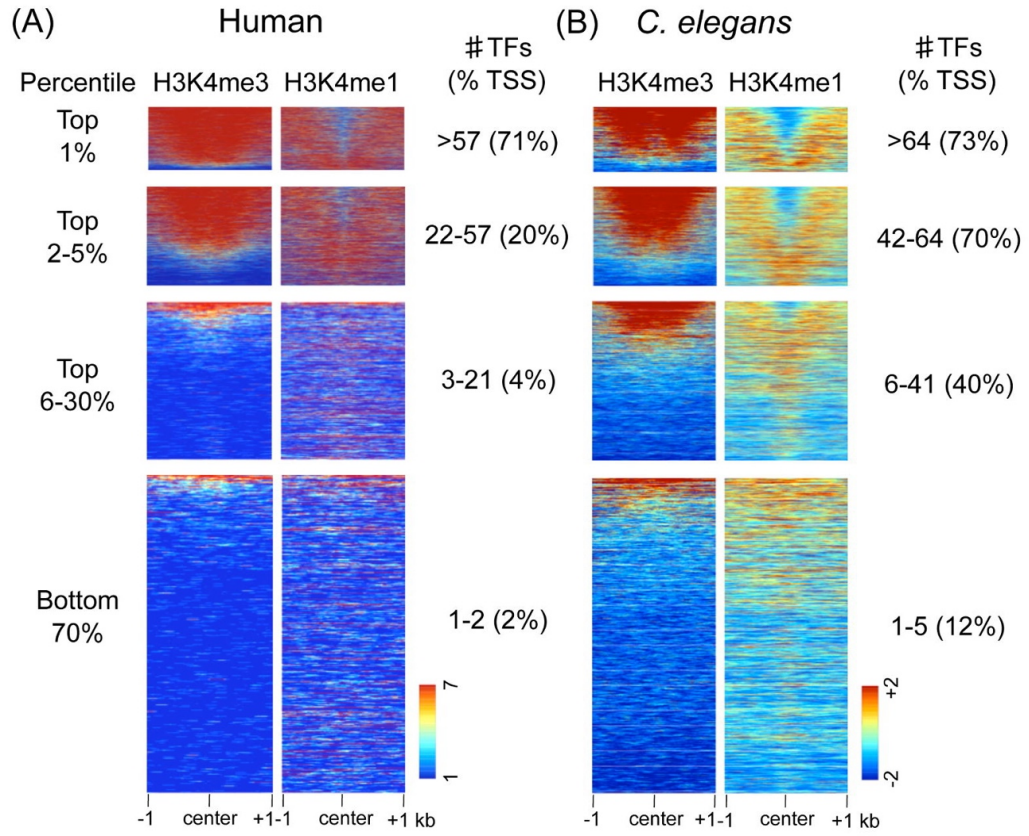


Figure 32 HOT regions exhibit promoter characteristics. H3K4me3 and H3K4me1 histone modifications shown around the midpoints of core HOT regions, ordered by the HOT score (number of unique TFs constituting the HTO regions). The heatmaps scales show input normalized linear (human) and log2 transformed (*C. elegans*) signal.

We first examined chromatin modifications at TF binding regions. I ordered regions by TF occupancy score (high to low) and investigated their association with H3K4me1 and H3K4me3 chromatin modifications. H3K4me1 is a known hallmark of enhancers, while H3K4me3 typically marks promoter regions - H3K4me1/ H3K4me3 ratio is often used as a proxy to determine if the given region is an enhancer or promoter (Robertson *et al.* 2008). In both organisms, highly occupied sites had chromatin marking that is characteristic of promoters – they were highly marked with H3K4me3 (**Figure 32**). This marking is particularly evident for the top 5% regions with highest factor occupancy. Highly occupied TF binding sites also usually overlapped an annotated TSS. The characteristics of HOT regions suggest that they are promoters. On the contrary, low-

occupancy regions showed low H4K4me3 and high H3K4me1 – characteristics of enhancers (**Figure 32**). In addition, whole genome analyses have shown similar results – in *C. elegans* 78% of genes downstream HOT regions were expressed in all tissues, as annotated using gene expression profiling (Spencer *et al.* 2011). This trend is even more evident for *H. sapiens,* where 91% of genes proximal to HOT regions were expressed in all cell types tested by ENCODE Project.

The finding that the HOT regions exhibit promoter characteristics was directly confirmed by my colleagues in reporter assay experiments. In this assay, Core HOT regions of 241–525bp and located at 150 bp to 4.7 kb upstream of the nearest annotated transcript start site were cloned directly upstream of a GFP reporter gene. All tested sequences drove GFP expression. In addition, most of 10 regions drove expression through all tissues in *C. elegans* (Chen *et al.* 2014a). Taken together, bioinformatics analyses and lab assay confirmed that the HOT regions are widely active, ubiquitous, core promoters.

3.3 HOT and COLD regions

I next wanted to further characterize differences between enhancer-like TFBS and HOT regions. As there was no formal definition of what HOT regions are, further than they should be highly occupied by TFs, I decided to coin two standard sets of extreme opposites of TFBS spectrum, named as “HOT” and “COLD” regions. I defined HOT regions as to 1% of TFBS ranked by number of unique factors binding, which translates to more than 64 factors binding in *C. elegans* and more than 57 factors in *H. sapiens*. COLD regions are low-occupancy sites where only a single factor was identified – this translates to 33.6% of regions in *C. elegans* and 30.8% in humans.

3.4 HOT regions are rich in CpG dinucleotides

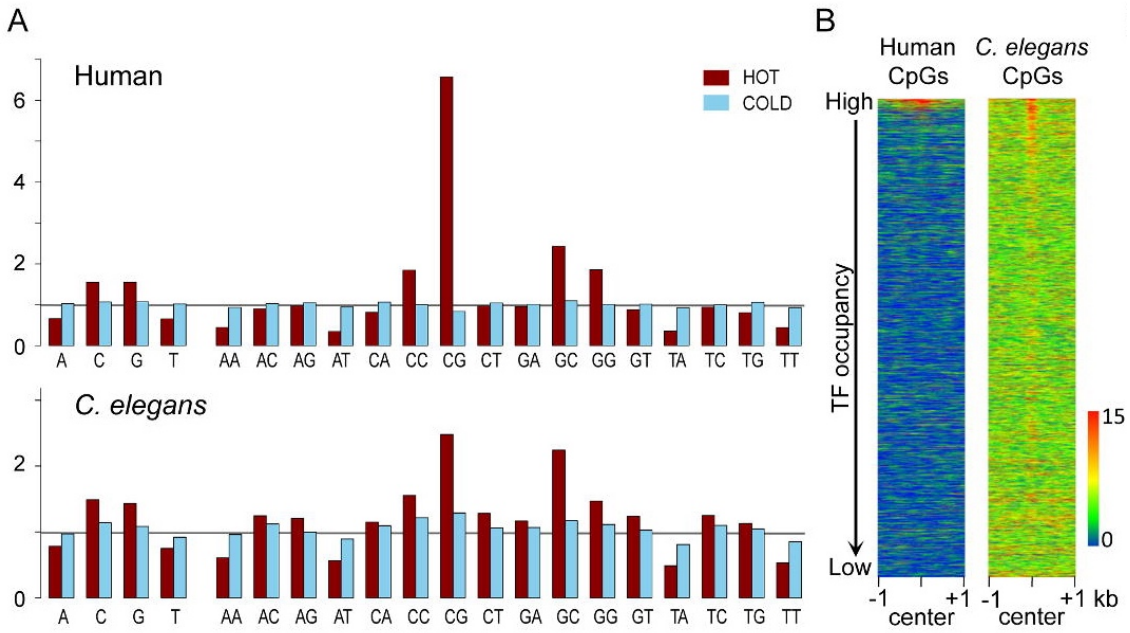


Figure 33 HOT regions are enriched for CpG. (A) The barplot shows the abundance of mono-nucleotides and di-nucleotides in HOT and COLD *C. elegans* and *H. sapiens* genomes. The y-axis shows nucleotide frequencies in assayed loci scaled to genome average – 1 represents no enrichment or depletion over genome average, numbers above 1 represent enrichment and numbers below 1 - depletion. In both species CpG is the most enriched dinucleotide in HOT regions, with weaker enrichment of GpC, GpG and CpC dinucleotides. (B) Heatmap showing CpG density on transcription factors overlapping sites. The sites are ordered by hotness score (number of unique TFs overlapping in given loci) – top of the heatmap represents HOT region, and bottom COLD regions. CpG content is calculated in 200bp moving window. Human CpG heatmap shows more regions in comparison to *C. elegans* heatmap, hence strong enrichment for CpG is visible only at very top of human CpG heatmap.

I first analysed sequence composition in these regions and observed that HOT regions had higher GC content than COLD regions (Figure 28). More strikingly, HOT regions in both *H. sapiens* and *C. elegans* exhibit high CpG dinucleotides concentration and show similar patterns of dinucleotide composition, with CpG being most highly enriched in both organisms (**Figure 33**). To further investigate the pattern of CpG dinucleotide enrichment I have plotted the 200bp moving average window CpG signal with SeqPlots (**Figure 34**). These analyses have revealed peaks of CpG enrichment around HOT regions – with the *C. elegans* peak being lower (peak height compared to background CpG enrichment), but much more concentrated than the *H. sapiens* peak.

This discovery is not surprising for human, where non-methylated CpG dinucleotides are known to be enriched at promoters (Gardiner-Garden & Frommer 1987; Illingworth & Bird 2009).

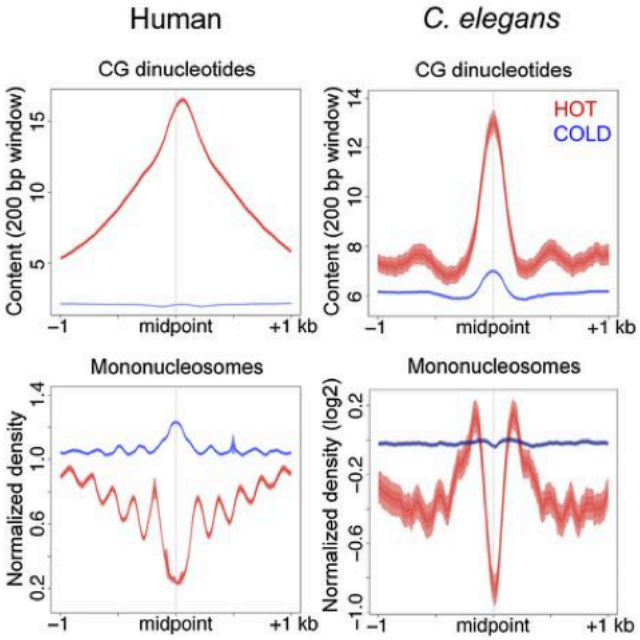


Figure 34 Hot regions are depleted for nucleosomes. Top panels: The distribution of CpG density around HOT (red) and COLD (blue) regions in human (left) and *C. elegans* (right) plotted with SeqPlots. Lines represents signal averages; darker ribbons represent standard error and lighter ones 95% confidence interval. The plotted region represents +/- 1kb from midpoint of HOT and COLD core regions. Bottom: The distribution of nucleosome density assayed by MNase data (see Materials and methods for details).

Since *C. elegans* does not have DNA methylation and classical, easily distinguishable CpG islands, it was unexpected to find CpG enrichment there. The CpG dinucleotides enrichment is not simply due to GC content – plotting CpG observed vs. expected ratio produces the same pattern of enrichment (**Figure 35**). These results suggest that CpG enrichment is a characteristic of active promoters in *C. elegans*, as is the case of *H. sapiens*.

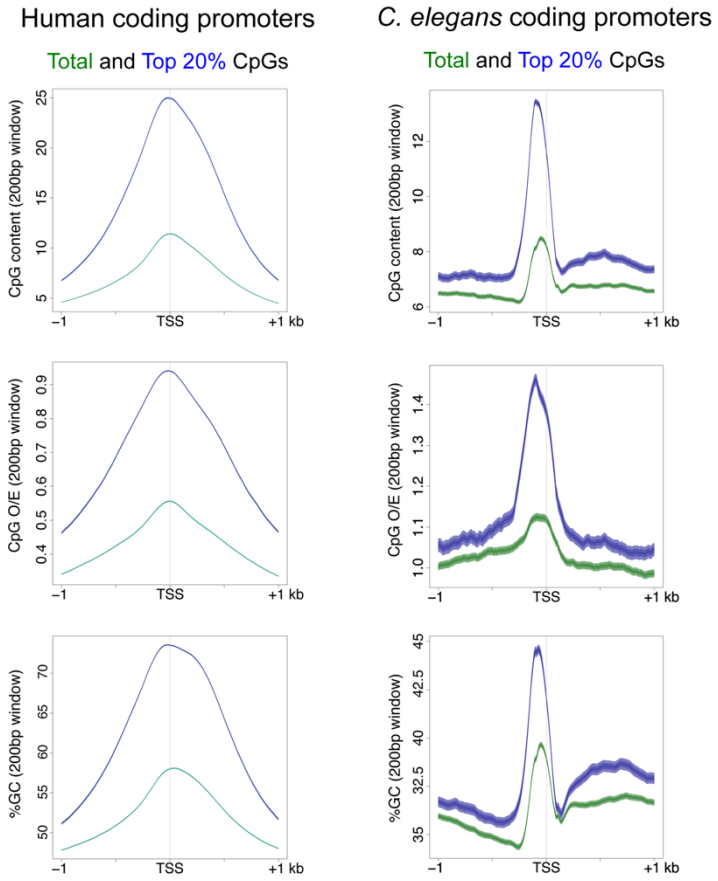


Figure 35 CpG dinucleotides and GC content distribution around *C. elegans* and human TSSs. Signal distributions for CpG density, normalized CpG (CpG O/E: observed/expected CpGs in 200bp window) and percentage of GC content (in 200bp) were plotted in 2 kb windows centred at all TSSs (green) or top 20% CpG TSSs (blue).

3.5 CpG dinucleotides are enriched on bulk of *C. elegans* and *H. sapiens* promoters

The previous analyses showed that CpG dinucleotides are enriched at HOT regions and that HOT regions are strong, ubiquitous promoters. I wanted to extend this analysis and ask if CpG enrichment is a general feature in both *C. elegans* and human promoters. Examining the CpG density and observed/expected CpG ratio showed clear CpG enrichment upstream of transcription start sites of coding genes in both species (**Figure 35**). Similarly to previous analyses, *C. elegans* CpG enrichment was lower, but more concentrated than human one (**Figure 35**). Also, in consensus with previous analyses, promoters exhibiting high CpG density tend to be ubiquitously expressed in comparison to CpG-poor promoters. Out of *C. elegans* promoters, I found that 56.3% of high CpG

and 21.8% of low CpG are ubiquitously expressed. This trend is even more evident in *H. sapiens,* where 68.6% of high CpG and 2.3% of low CpG are promoters of ubiquitously expressed genes. In conclusion, CpG dinucleotides are enriched on widely active promoters.

To better comprehend this result, it is worth to highlight the difference of promoter architecture and function between *C. elegans* and human in the light of discovery that *C. elegans* active promoters are also CpG rich. In *H. sapiens* in CpG islands of active promoters the cytosines are typically non-methylated, whereas non-promoter or inactive CpGs are usually cytosine methylated. Also, non-regulatory regions of the human genome are typically CpG depleted. DNA methylation dynamics is regarded to regulate both the recognition of CpG dinucleotides by expression machinery and change the mutation rate of cytosine, which drives global depletion of unprotected CpGs in human genome. In the *C. elegans* genome neither global CpG depletion nor DNA methylation are present. These differences might be related to the different patterns of CpG enrichment on promoters in these two species – human shows much stronger and wider CpG enrichment, while CpGs in *C. elegans* are better positioned as a sharp peak in close proximity to TSS.

3.6 Nucleosome depletion in promoters is associated with CpG density independently of expression and GC content

In addition to CpG enrichment, I also found that HOT regions are nucleosome depleted in both organisms, compared to COLD regions (**Figure 36**). I next investigated if nucleosome depletion in HOT regions is driven by their CpG density or expression activity.

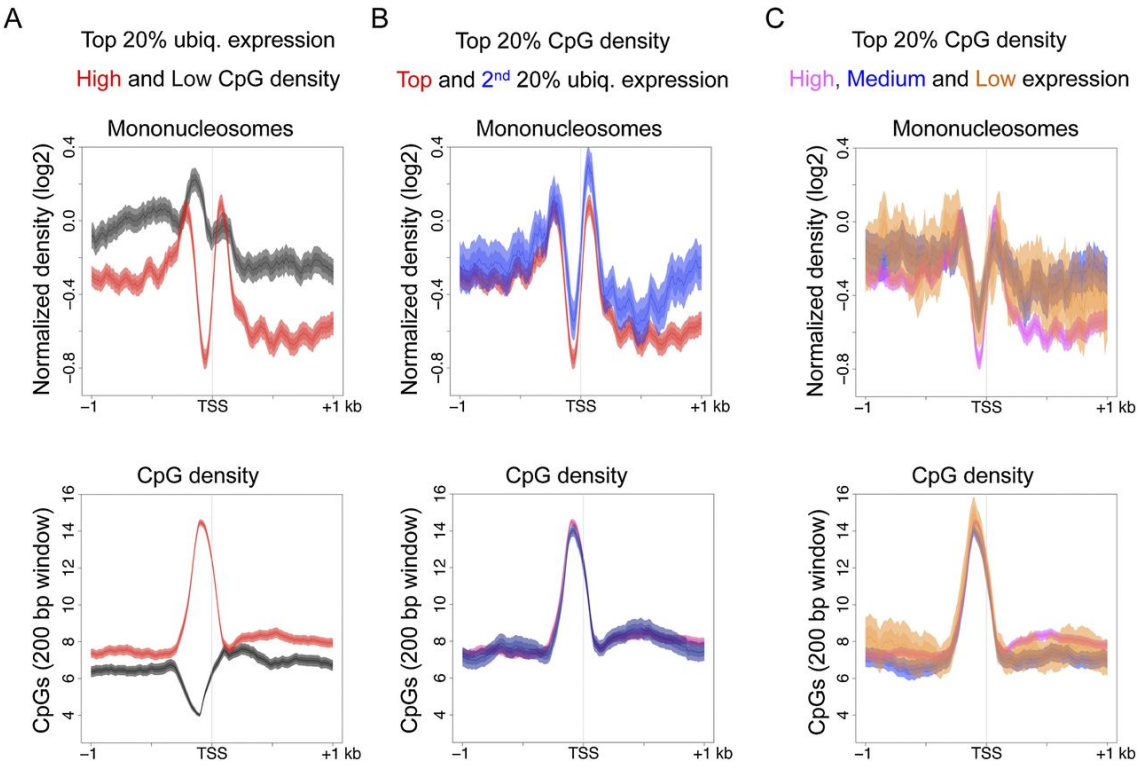


Figure 36 High CpG content promoters show nucleosome depletion in *C. elegans*. (A) Mononucleosome density (top) and CpG density (bottom) plotted in top 20% of promoters ranked by expression and divided into high (top 20% CpG, red) and low (bottom 20% CpG black) CpG content. (B) Mononucleosome density (top) and CpG density (bottom) plotted in top 20% of ubiquitously active promoters ranked by CpG density and divided into highly (top 20% expression, red) and low (bottom 20% expression, blue) expressed genes. (C) Mononucleosome density (top) and CpG density (bottom) plotted in top 20% of promoters ranked by CpG density and divided into three bins based on gene activity – highly active (top 20% expression, pink), active (middle 20% expression, blue), and lowly expressed/inactive (bottom 40% expression, orange).

It was previously shown that mammalian CpG rich regions exhibit strong nucleosome depletion which is independent of RNA polymerase activity (Fenouil *et al.* 2012; Vavouri & Lehner 2012). Therefore, it was of interest to investigate if *C. elegans* CpG enriched regions show the same trend. To do this I selected promoters of the top 20% of expression of ubiquitously active *C. elegans* genes and then defined high and low CpG content groups by selecting top and bottom 20% of CpG density. I then compared nucleosome occupancy between high and low CpG content of ubiquitously active promoters. I observed there is a striking difference in nucleosome pattern – highly active promoters with high CpG density show strong nucleosome depletion, whereas highly active promoters with low CpG density show no nucleosome depletion (**Figure**

36A). This result stays unchanged when normalising CpG density to GC content using observed over expected (O/E) CpG ratio. This data suggests the CpG density, not the expression activity is associated with nucleosome depleted regions in promoters. To strengthen this result, I also analysed GC content in high and low CpG regions.

As mentioned before, CpG related nucleosome depletion stayed valid when analysing O/E CpG ratio, however the nucleosome depletion could still be primarily driven by GC content just showing even stronger effect for CpG reach regions. I observed that both low and high CpG density regions in highly active promotors exhibit similar level of GC enrichment (**Figure 37**B and C). This results strongly underlines that it is CpG density, not a GC content or expression activity, which is associated with nucleosome depletion on promoters (**Figure 37**).

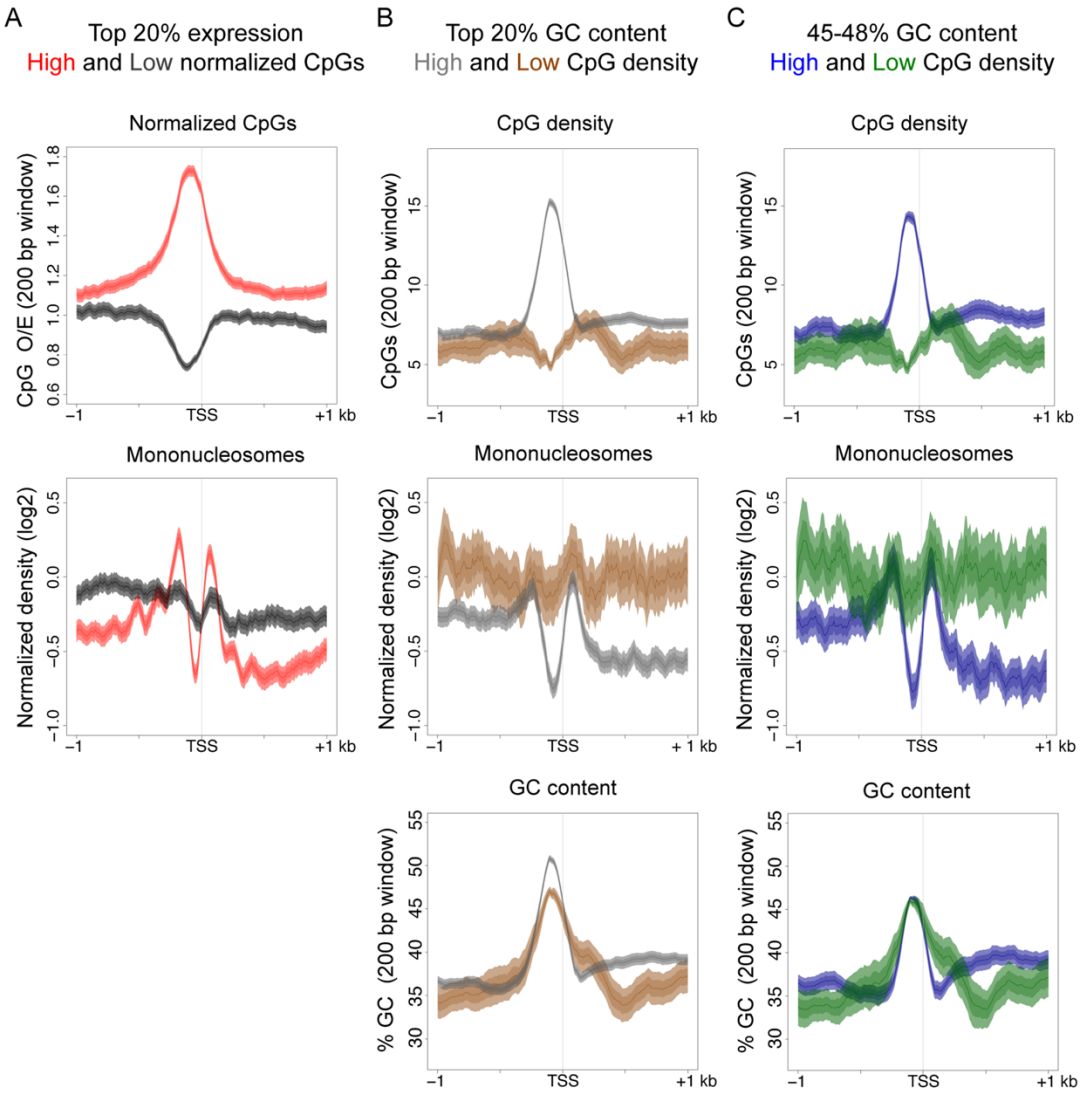


Figure 37 Nucleosome depletion is linked with high normalized CpG density but not high GC content at *C. elegans* promoters. (A) Plots of normalized CpG (O/E: observed/expected) and nucleosome density across promoters of ubiquitous genes in the top 20% of expression with high (red, top 20%) or low (dark grey, bottom 40%) normalized CpG content (O/E: CpGs in a 200bp window). (B) Signal distributions of ubiquitous promoters in the top 20% of GC content separated into those in the top 20% (light grey) or bottom 40% (brown) of CpG content. (C) The indicated signal distributions of ubiquitous promoters with a narrow range of closely matched high GC content (45-48%) in the top 20% (blue) or bottom 40% (green) of CpG content. CpG and GC contents were calculated in windows 200 bp upstream of TSSs.

3.7 In high CpG, ubiquitously active promoters transcriptional activity shows weak correlation with nucleosome density

In general, the nucleosome density at promoters is inversely proportional to activity,

meaning the promotes of highly expressed genes tend to be nucleosome depleted. In

previous results I have shown that in top 20 percentile of expression the CpG density is more deterministic for nucleosome density than gene expression. To asses if CpG is a better predictor for accessibility in general I have analysed all highly expressed and ubiquitously active promoters. In general, ubiquitously active genes are highly expresses, with 94% of them falling into top 40 percentile of protein coding gene expression. In the first instance I than separated top 20% of CpG genes into top and bottom 20th percentile of expression. In both groups I observed a similar depletion of nucleosomes, concluding that in CpG rich, ubiquitous promoters the level of transcriptional activity has little to no relationship with nucleosome occupancy (**Figure 36**B).

3.8 In all coding genes both CpG content and RNA polymerase activity contribute to accessibility

Next, I have divided all high (top 20%) CpG promoters into expression bands – now the genes in lower expression bands are enriched for tissue specific promoters, while higher band represents previously analysed ubiquitous, housekeeping genes. Promoters across all bands of expression exhibit clear nucleosome depletion, however the level of depletion is proportional to expression activity, meaning the more expressed the band is, the higher and more uniform (lower standard error) the nucleosome depletion is (**Figure 36**C) (Chen *et al.* 2014b). This observation suggests that CpG density is a strong determinant of chromatin accessibility, but alone it is not sufficient for creating open chromatin state and some additional regulation determines the nucleosome depletion level.

3.9 C. elegans CXXC protein CFP-1 is targeted to CpG-rich promoters

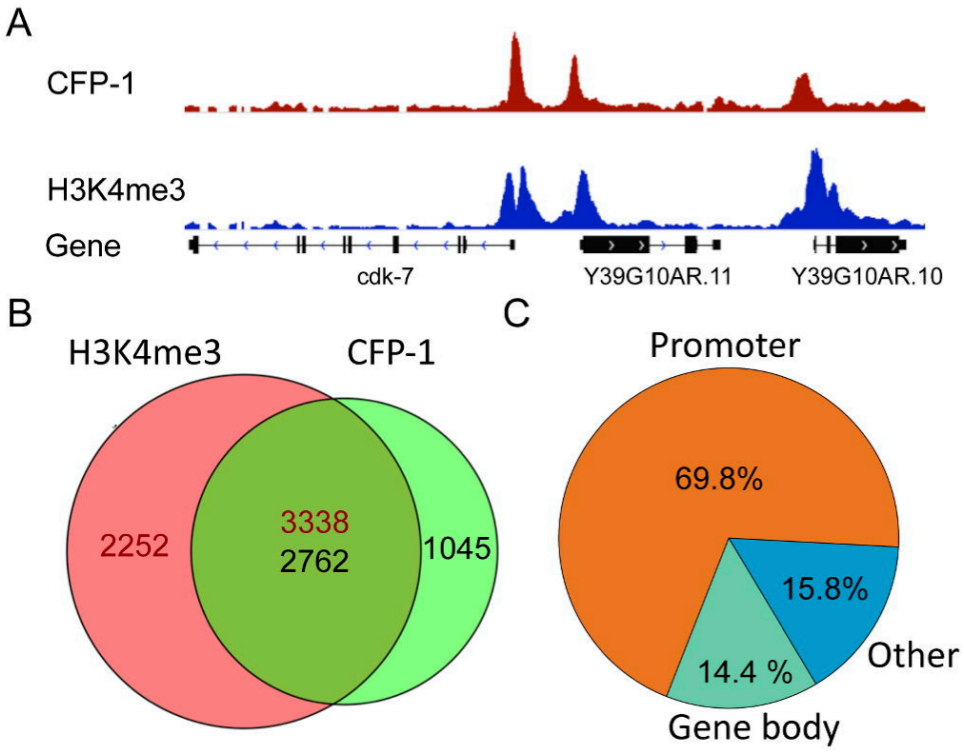


Figure 38 CXXC protein CFP-1 is targeted to H3K4me3 marked, CpG-rich promoters. (A) genome browser view showing a typical pattern of H3K4m3 and CFP-1 co-occurrence on coding gene promoters. (B) Venn diagram representing overlaps between CPP-1 and H3K4me3 peak calls (see methods for more). In the middle of Venn plot red number represents H3K4me3 in peaks CFP-1 – H3K4me3 overlapping regions, while black number represents CFP-1 peaks in the same region. The number of H3K4me3 peaks is larger than CFP-1 peaks, because H3K4me3 often exhibits double peak pattern – in such regions there is a single CFP-1 peak and two H3K4me3 peaks annotated by peak caller software and overlapping each other. (C) Genomic annotation of CFP-1 peaks. Promoters (shown in orange) are defined as +/- 500bp region from coding TSS, as defined in CAP promoters set based on Chen et al. 2013 and Kruesi et al. 2013 (see methods for more). When assigning peaks, for peaks overlapping more than one annotation class the priority was given to promoters, than to gene body, and peaks not overlapping with any of the above were classified as “Other”

CXXC-type zinc finger protein 1 (CXXC1) is a protein that shows specific binding to unmethylated CpG islands in higher eukaryotes, with a specific preference for a CpGG motif (Xu *et al.* 2011). CXXC1 binds to CpG dinucleotides through its CXXC domain and is a member of the COMPASS/SETD1 complex, which trimethylates lysine 4 of histone H3 (Lee & Skalnik 2005; Tate *et al.* 2010; Thomson *et al.* 2010). Despite the lack of DNA methylation (Simpson *et al.* 1986) and unmethylated CpG islands in *C.*

elegans there is a CXXC1 ortholog with a conserved CXXC domain called CFP-1. Furthermore, previous reports demonstrated that CFP-1 is required for global H3K4 methylation (Li & Kelly 2011; Simonet *et al.* 2007). Led by these findings we wanted to investigate if, similarly to *H. sapiens* CXXC protein, CFP-1 bound to CpG rich promoter regions in *C. elegans*.

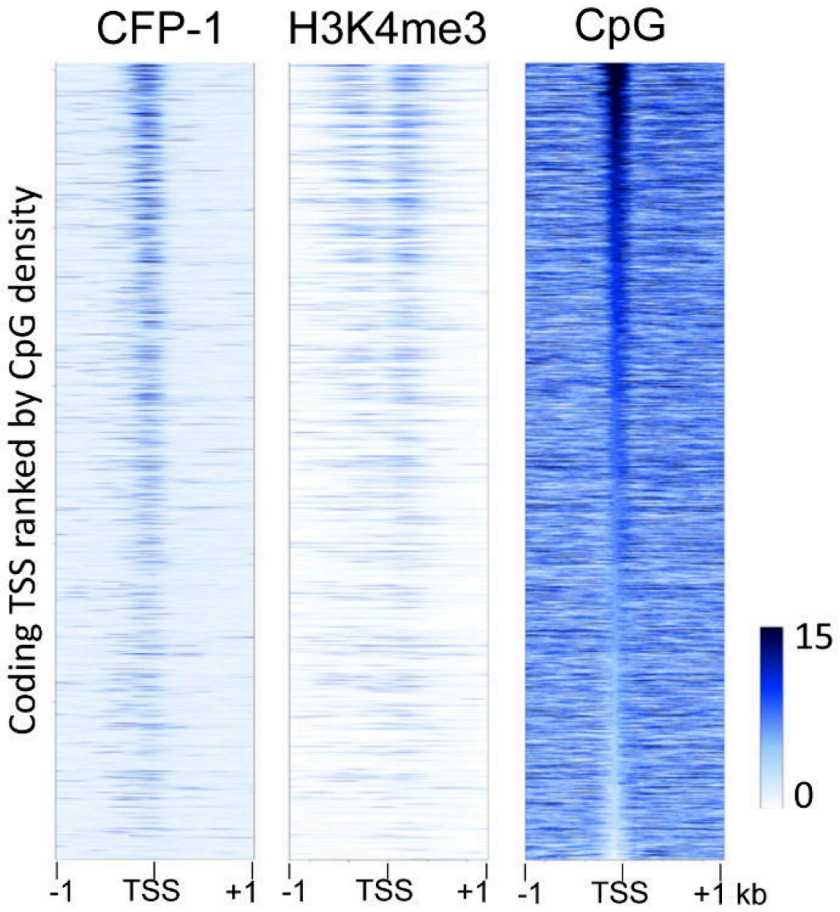


Figure 39 CFP-1, H3K4me3 and CpG enriched regions overlap with each other. The heatmaps show profiles of CFP-1, H3K4me3 ChIP-seq signals and CpG density in 200bp moving window plotted in +/- 1kb from coding TSSs identified in Chen et al. (2013). The plotted regions are ordered by CpG content.

To investigate the binding pattern of CFP-1, my colleagues made a transgenic strain that expresses GFP-tagged CFP-1 in addition to the endogenous CFP-1 and carried out ChIP-seq experiments, and I analysed the data. I found that most CFP-1 binding sites are located on promoters and overlap regions strongly marked with H3K4me3 (**Figure 38**A). I also observed that CFP-1 is enriched at CpG rich sites, and that CpG distribution narrowly peaks over CFP-1 peak calls (**Figure 39**). More precisely, CFP-1

peak summits reside in between -1 and +1 nucleosome marked by H3K4me3, overlapping the middle of the nucleosome depleted region (NDR) with very narrow peak distribution. Strikingly, there is a very strong correlation between CpG density, CFP-1 binding and H3K4me3 marking (**Figure 39**). Furthermore CFP-1 peaks that do not map upstream of annotated TSS exhibit chromatin signatures similar to promoters – high H3K4me3 to H3K4me1 ratio and nucleosome depletion, which suggests they might be promoters of non-coding or not annotated genes (**Figure 40**A). CFP-1 binding is strongly associated with HOT, and not associated with COLD regions that exhibit enhancer characteristics (**Figure 40**B).

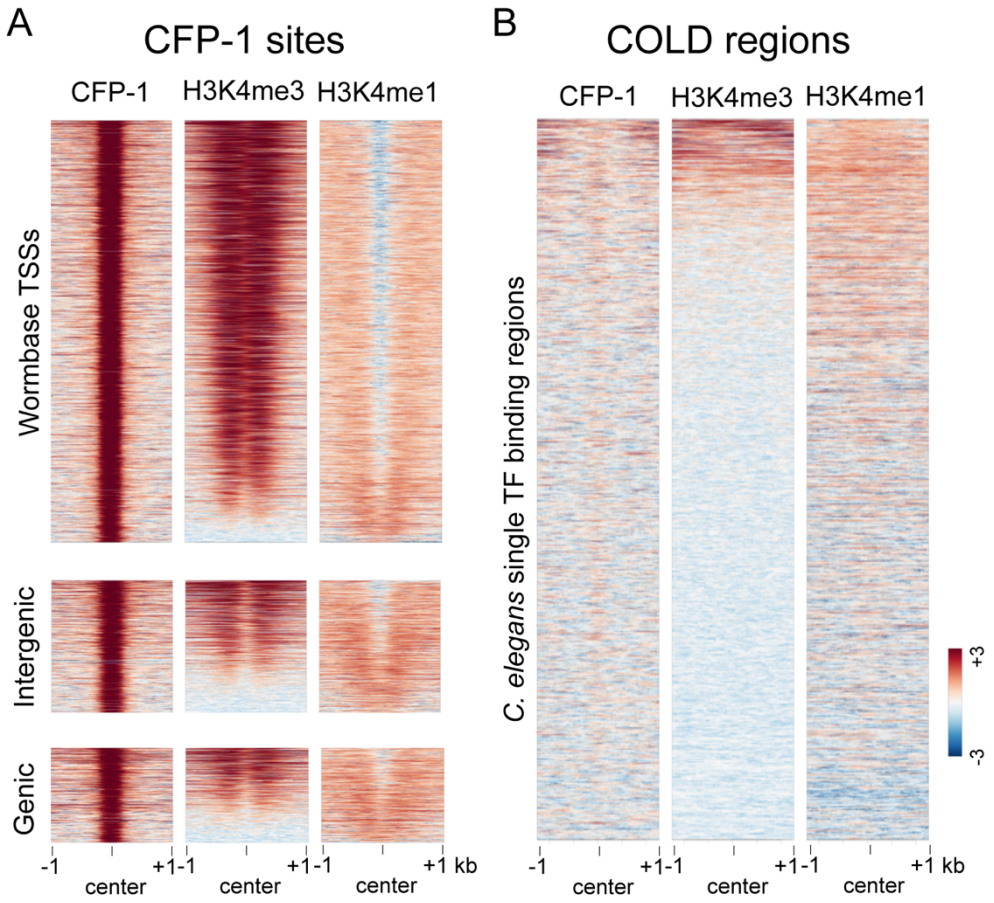


Figure 40 Chromatin signatures at CFP-1 sites. Heat map plots of CFP-1, H3K4me3, and H3K4me1 ChIP-seq signal across the indicated regions. (A) CFP-1 binding sites separated by genomic location: Wormbase TSSs, intergenic regions, and genic regions (within genes). CFP-1 sites mapping to intergenic and genic regions are enriched for having promoter-like chromatin signatures (H3K4me3high and H3K4me1low), similar to those at Wormbase TSSs (those mapping +/- 500 bp of a Wormbase transcript start site), (B) In contrast, CFP-1 has low signal in COLD regions, which mainly display

enhancer-like chromatin features (H3K4me3low and H3K4me1high). All data sets are plotted in 1kb windows centered at the midpoint of core regions. The scale gives the enrichment of ChIP signals in log2.

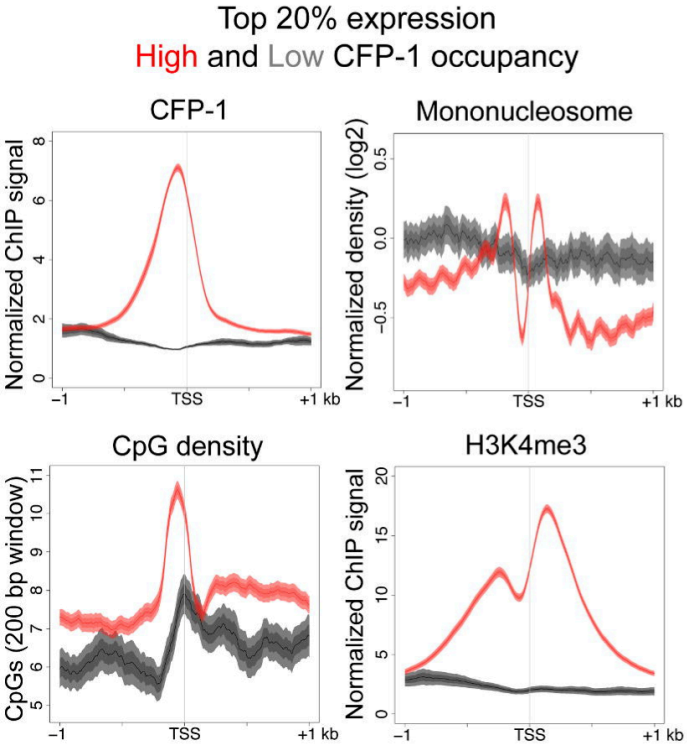


Figure 41 Highly expressed promoters bound by CFP-1 show nucleosome depletion, CpG density peak and strong H3K4me3 marking. Plotted regions show +/- 1kb from TSS of highly active genes (top 20% expression) divided in regions with high CFP-1 marking (top 20% ChIP signal, red) and low CFP-1 marking (bottom 20% ChIP signal, blue).

Moreover, highly expressed genes (top 20 percentile of expression) with high CFP-1 (top 20 percentile of binding locations) also show high CpG content, nucleosome depletion and H3K4me3 marking, but highly expressed genes with little to no CFP-1 (bottom 20 percentile) does not exhibit such characteristics (**Figure 41**). That suggests that CFP-1 binding may be important for nucleosome depletion, but not for high expression. Similarly to CpG density analyses, all promoters in the top 40th percentile of expression and with high CFP-1 binding also show H3K4me3 marking and nucleosome depletion (**Figure 42**). This result shows that binding to CpG-rich regions that are H3K4me3-marked and nucleosome depleted is conserved between *H. sapiens*

CXXC1 and *C. elegans* CFP-1 orthologs. Either Cfp1 or the CpG density might drive the nucleosome depletion.

.

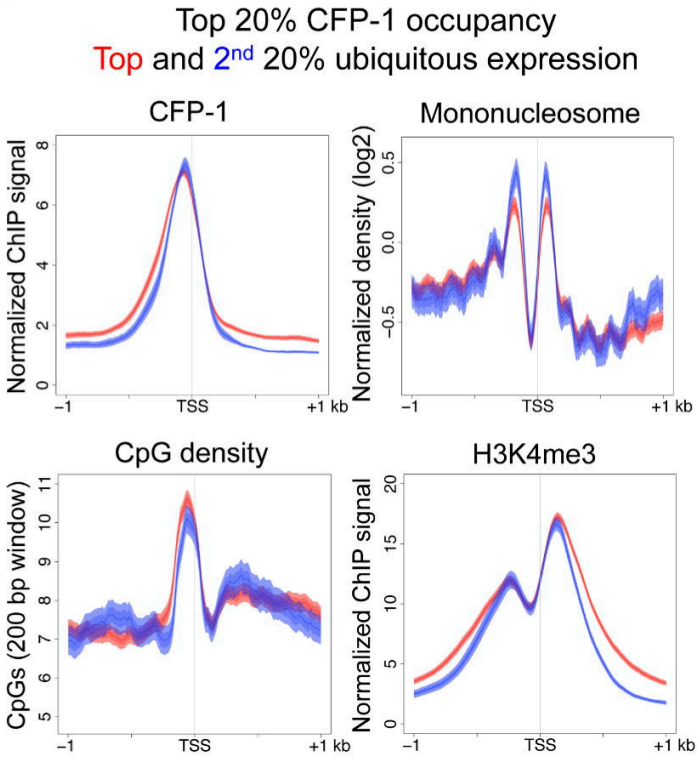


Figure 42 Regions bound with CFP-1 show strong promoter characteristics – nucleosome depletion, CpG density peak and strong marking with H3K4me3 in 1st and 2nd pentile of ubiquitous expression. Plotted regions show +/- 1kb from TSS of ubiquitously expressed genes, highly marked by CFP-1 - top pentile of CFP-1 ChIP signal. Further, these regions are divided based on expression measurements into 1st (red) and 2nd (blue) pentile of ubiquitous expression.

3.10 CFP-1 and SET-2 are required for H3K4me3 deposition on promoter regions

Previous analyses showed that CFP-1 co-localises with the H3K4me3 histone modification. It was also reported that CFP-1 is a part of COMPASS complex in *S. cerevisiae* and SET1 complex in *H. sapiens* – the homologous complexes that contain SET domain methyltransferase and are required for H3K4me3 deposition (Lee & Skalnik 2005). Finally, it was also reported in *C. elegans* that RNAi of *cfp-1* causes global loss of H3K4me3 (Li & Kelly 2011).

In order to confirm this result and establish if the loss of H3K4me3 in *C. elegans* is specific for CFP-1 binding loci I analysed H3K4me3 ChIP-seq experiments done by my colleagues in *cfp-1* and *set-2* (encoding the *C. elegans* ortholog of SET1) loss of function mutants. Because it was known that H3K4me3 levels are very low in the mutants, *C. briggsae* chromatin was included to control for ChIP efficiency. The data were then, normalised based on the signal in *C. briggsae.* Using the internal control, we could accurately quantify the loss of H3K4me3 in *cfp-1* and *set-2* mutant backgrounds. The spike-in normalization was done by my colleague Ni Huang. As for April 2018 we have not performed SET-2 ChIP-seq and there are no SET-2 ChIP-seq experiments in Gene Expression Omnibus (GEO) repository.

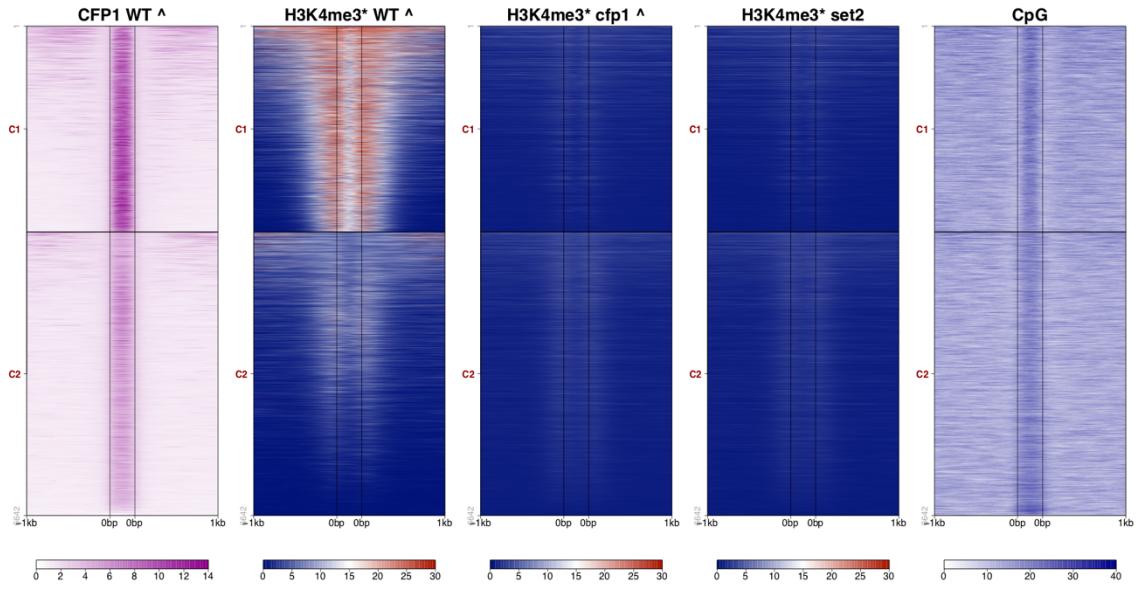


Figure 43 The heatmap showing depletion of H3K4me3 on CFP-1 peak sites in *cfp-1* and *set-2* mutant backgrounds. CFP-1 peak calls were divided into 2 distinct clusters for further analyses (1) Cluster 1: high CFP-1/total loss of H3K4me3; (2) Cluster 2: low CFP-1 sites/low H3K4me3. "^" denotes tracks used for clustering and "*" denotes tracks normalized with *C. briggsae* chromatin spike-in method. Last heatmap shows CpG di-nucleotide density in 200bp window.

I observed a significant loss of H3K4me3 in both mutant backgrounds, with only residual marking left in certain loci in cluster C2. I theoreticized, that the strong loss of H3K4me3 is primary on promoters, while the weaker loss is more specific for enhancers. To check this hypothesis I plotted H3K4me3 in WT and mutant

backgrounds, along CFP-1 on promoters and enhancers as defined in (Janes *et al.* 2018). Indeed, there was much stronger marking of H3K4me3 on promoters in WT, compared to enhancers, and nearly complete loss of H3K4me3 in *cfp-1* and *set-2*. However, the loss of H3K4me3 was also very strong in the enhancer cluster: much stronger than at some loci shown in LoConf CFP-1 sites (C2) (**Figure 43**). HiConf CFP-1 sites are very strongly enriched for promoters - 77.69% overlap promoters, while 16.65% overlap enhancers. C2 is more equally divided, with 40.89% overlapping promoters and 35.88% overlapping enhancers. Interestingly, while nearly all sites in C1 overlap either promoter or enhancer (94.34%), in C2 this overlap is much weaker (76.77%), leaving 23.23% sites not overlapping with an annotated regulatory element. The data suggest that the most of retained H3K4me3 signal cannot be attributed to annotated regulatory elements.

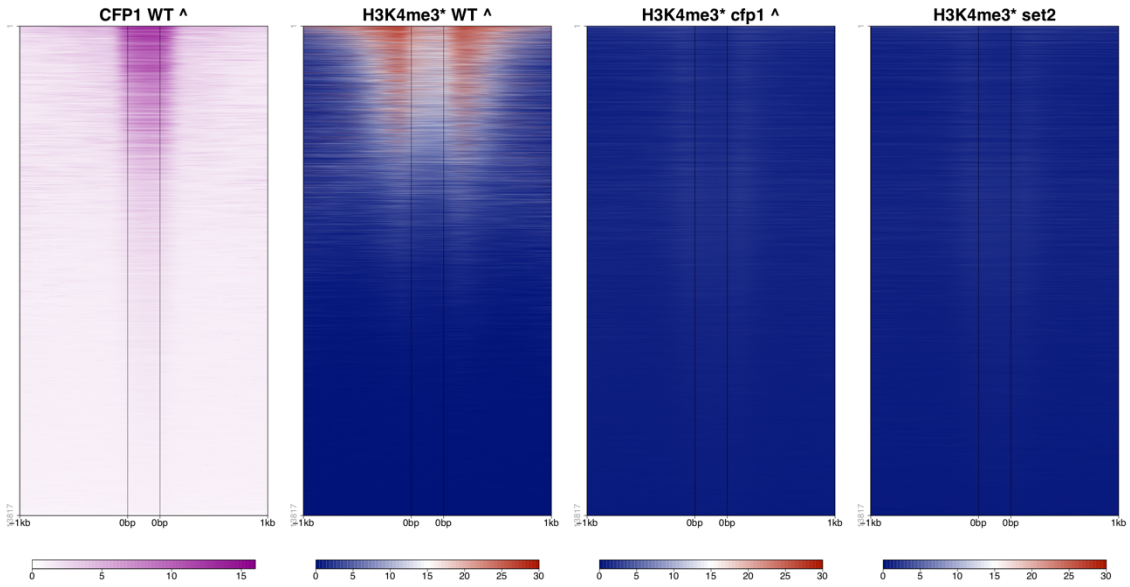


Figure 44 Loss of H3K4me3 on promoters. H3K4me3 in WT, *cfp-1* and *set-2* backgrounds, along CFP-1 on promoters as defined in (Janes *et al.* 2018)

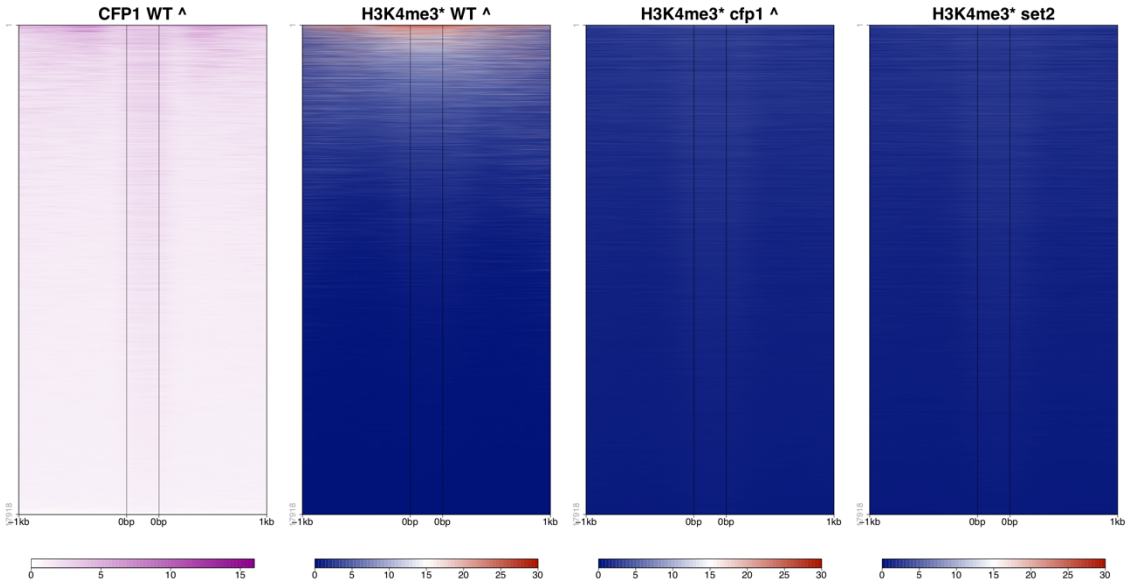


Figure 45 Loss of H3K4me3 on enhancers. H3K4me3 in WT, *cfp-1* and *set-2* backgrounds, along CFP-1 on enhancers as defined in (Janes et al. 2018)

Furthermore, the patterns in *cfp-1* and *set-2* loss of function are remarkably similar, supporting the hypothesis that CFP-1 works upstream of SET-2, recruiting it to the particular, mostly promoter, loci on DNA. I then focussed my attention to CFP-1 binding loci, as identified by peak calls in CFP-1 wild-type ChIP. I wanted to focus on a stringent set of peaks, so after calling peaks separately on each replicate with MACS2 (Zhang *et al.* 2008) (adjusted $p<0.01$), I produced a final set by intersecting these individual peak calls.

The heatmap plot on CFP-1 sites shows a very strong loss of H3K4me3 in both *cfp-1* and *set-2* mutants (**Figure 43**). I found the sites could be clustered into two classes of CFP-1 binding/loss of H3K3me3 – cluster 1 (C1) where there is strong CFP-1 binding and almost complete loss of H3K4me3 signature, and cluster 2 (C2), where CFP-1 binding is weaker in comparison to cluster one, and the loss of H3K4me3 is weak (**Figure 46**). This pattern suggests that CFP-1 may have different modes of action. Since no antibody against CFP-1 was available, CFP-1 sites were mapped using a transgenic GFP – CFP-fusion construct that was integrated into genome at MosSCI insertion site under a ubiquitous promoter. A potential caveat to the binding data is that

the additional dosage of CFP-1 protein might cause additional binding sites that are normally not strongly bound by endogenous CFP-1. However, the loss of H3K4me3 in *cfp-1* mutants at these sites suggests that they are true binding regions. In further analyses I will use this clustering to distinguish between two modes of CFP-1 binding – the high confidence SET-2/COMAPSS CFP-1 sites (C1, 2793 sites) and low confidence sites (C2, 3849 sites).

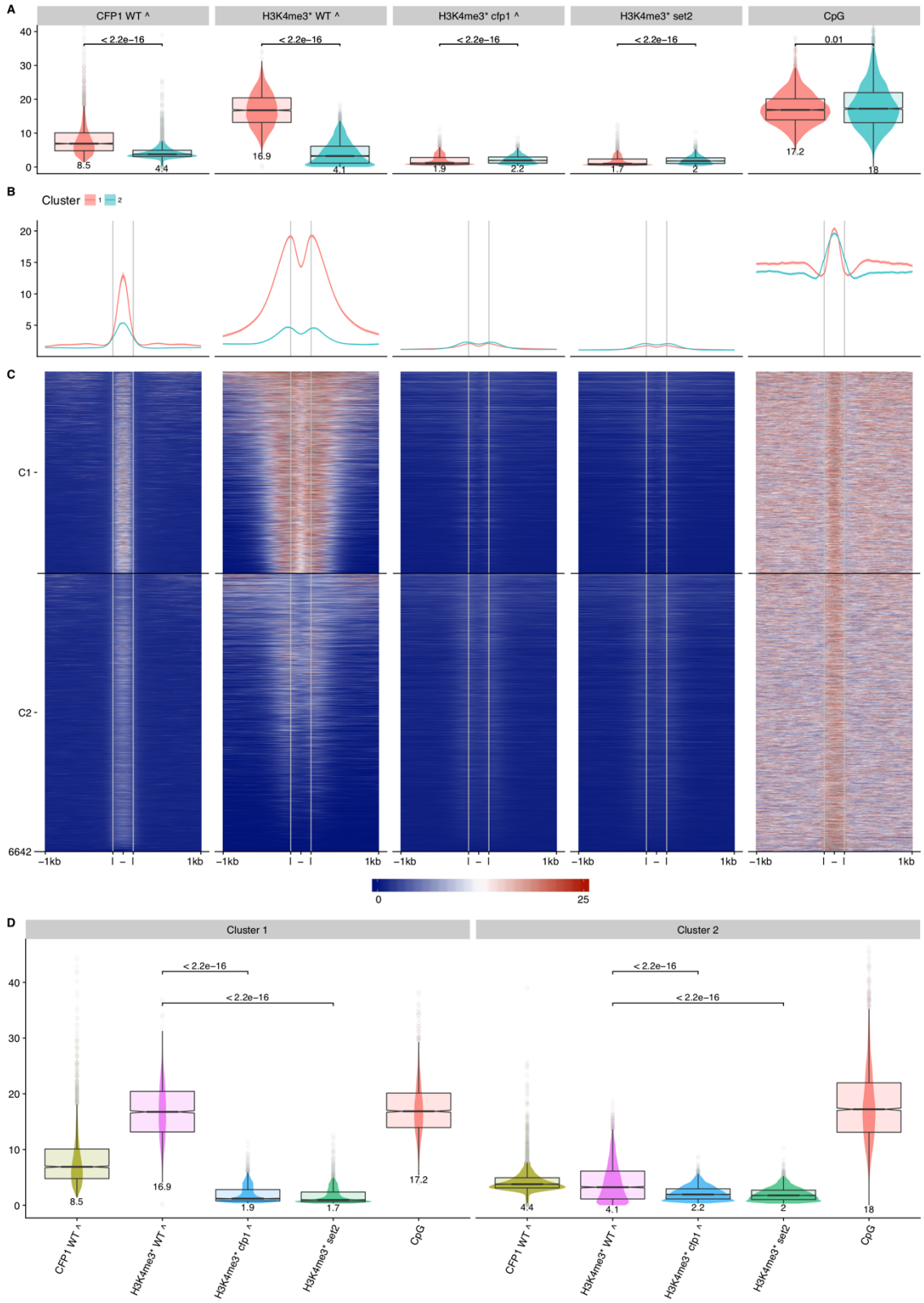


Figure 46 Combined plot showing the quantification of depletion of H3K4me3 and CpG di-nucleotide density in 200bp window on CFP-1 peak sites in *cfp-1* and *set-2* mutant backgrounds. CFP-1 peak calls were divided into 2 distinct clusters for further analyses (1) Cluster 1: high CFP-1/total loss of H3K4me3; (2) Cluster 2: low CFP-1 sites/low H3K4me3. "^" denotes tracks used for clustering and "*" denotes tracks normalized to *C. briggsae* chromatin using the spike-in method. Panels A and D show U-test p-values. The numbers below box plots represent the mean of quantified signal.

In order to further strengthen these observations, I quantified the signal at CFP-1 peak loci and plotted it using boxplots and performed Mann-Whitney U-test to establish if there was a difference in signal distributions between WT and mean backgrounds and between clusters C1 and C2. The quantification shows there is a clear difference (**Figure 46**A) between C1 and C2 in wild-type CFP-1 and H3K4me3. For *cfp-1* and *set-2* mutant backgrounds there is a very significant loss of H3K4me3 compared to wild-type in both clusters (**Figure 46**D). However, the residual H3K4me3 signal quantification in C1 and C2 is opposite to WT – while C1 showed significantly larger marking than C2 in WT, in mutant backgrounds the leftover signal in C2 was significantly higher than in C1 (**Figure 46**A).

To investigate differences between the clusters, I plotted CpG density. I found that both clusters have a strong CpG signature in the peak region (**Figure 43**). However, I observed that CpG sites in C1 have a tighter distribution in comparison to C2 (**Figure 46**B and C, leftmost panel). Also, in C1 the background levels of CpG are higher, and there is a characteristic double valley pattern, co-localising with double peak pattern of H3K4me3. This might indicate that locations of -1 and +1 histones from strong CFP-1 binding sites are depleted for CpG di-nucleotides.

3.11 H3K4me3 shows different modes of enrichment in CFP-1 peak regions

To investigate the CpG patterns further, I separated the CFP-1 sites into a larger number of clusters to see if I could find a higher complexity structure. I found that six clusters could pull out additional patterns (**Figure 47**).

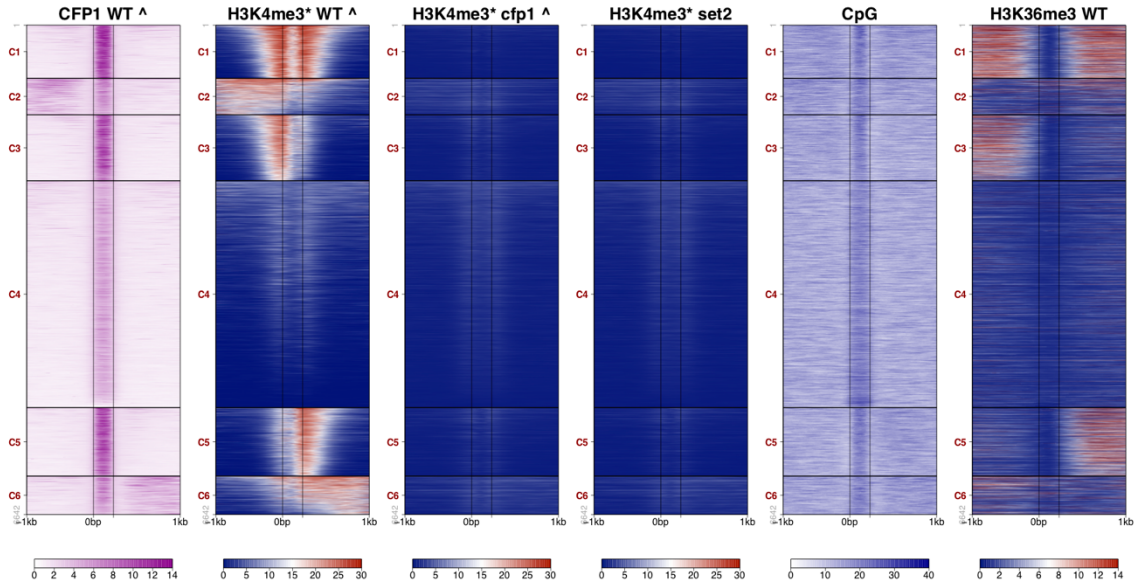


Figure 47 Heatmap showing depletion of H3K4me3 on CFP-1 peak sites in *cfp-1* and *set-2* mutant backgrounds. CFP-1 peak calls were divided into 6 clusters, showing diverse pattern of H3K4me3 marking "^" denotes tracks used for clustering and "*" denotes tracks normalized with *C. briggsae* chromatin spike-in method. 5th heatmap shows CpG di-nucleotide density in 200bp window. Last heatmap panel shows H3K36me3 pattern in wildtype. H3K36me3 indicates directionality of expression.

This setup revealed an interesting structure of H3K4me3 marking. Cluster (C1) has the strongest CFP-1 signal and H3K4me3 equally marks the -1 and +1 nucleosomes. These are bidirectional promoters as shown by the H3K36me3 pattern, which marks transcription elongation (Kolasinska-Zwierz *et al.* 2009) (**Figure 47**). This cluster also exhibits strongest loss of H3K4me3 marking in both *cfp-1* and *set-2* backgrounds - confirmed by quantifications and linear plots (**Figure 49**A, B and D). Clusters C3 and C5 are unidirectional promoters in which H3K4me3 marks the +1 nucleosome, and clusters C2 and C6 appear to be weak CFP-1 sites that are near strongly marked H3K4me3 sites. Cluster C4 shows weaker binding of CFP-1 than other clusters and little or no H3K36me3 in either direction suggesting the sites are either weak promoters or are not promoters. In both mutant backgrounds there is a smaller loss of H3K4me3 binding than in other clusters. However, this cluster still shows good enrichment of CpG, and similarly to cluster C2 in previous analyses this cluster has a wider distribution of CpG around CFP-1 peak than other clusters (**Figure 49**B).

3.12 H3K4me3 depletion and CpG patterns show correlation of directionality of expression

Intriguingly, each of the clusters exhibits a unique pattern of CpG enrichment: there is a distinctive dip of CpG enrichment that co-localises with nucleosomes not marked with H3K4me3 – downstream of CFP-1 peak in C3, and upstream in C5. Similarly to C1, there is also a sharp peak and high background of CpGs around nucleosome depleted region (**Figure 48**, **Figure 49**).

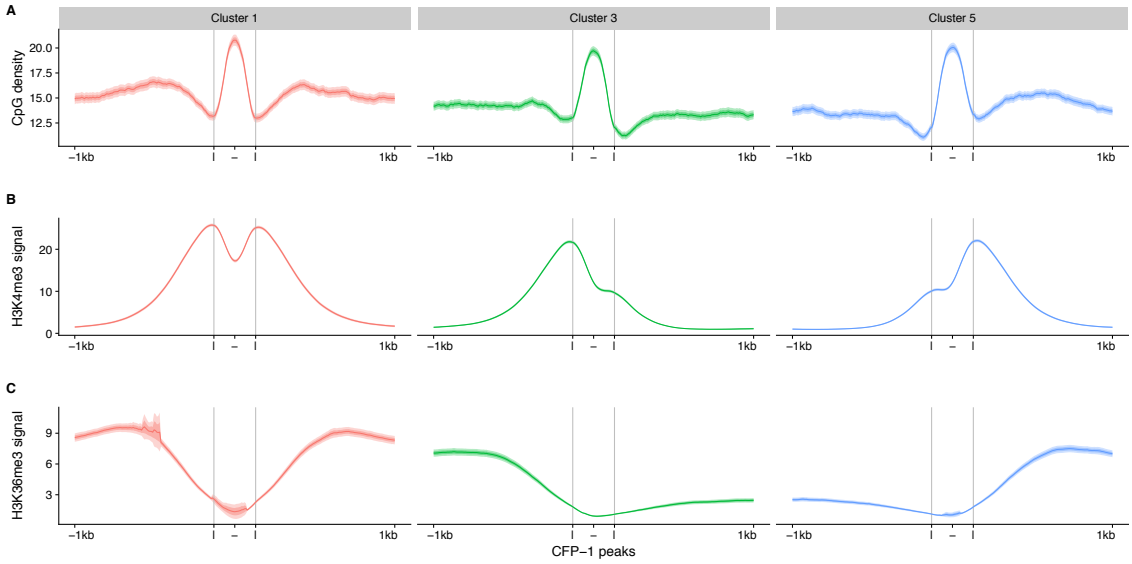


Figure 48 Profile plots showing individual CpG profiles around CFP-1 binding loci in clusters 1, 3 and 5 (C1, C3, C5) (A), alongside H3K4me3 spike-in normalised signal (B) and H3K36me3 rBEADS normalised signal (C). Corresponding heatmaps are shown on **Figure 49**

Cluster 1 is a bidirectional expression cluster - -1 and +1 nucleosomes are symmetrically marked with H3K9me3, and there is a symmetric enrichment of H3K36me3 on both sides of nucleosome depleted CFP-1 binding site. Cluster 3 is reverse strand monodirectional promoter – there is a stronger, asymmetric depletion of CpG density downstream CFP-1 binding site and strong H3K4me3 marking of -1 nucleosome, followed by asymmetric upstream marking of H3K36me3.

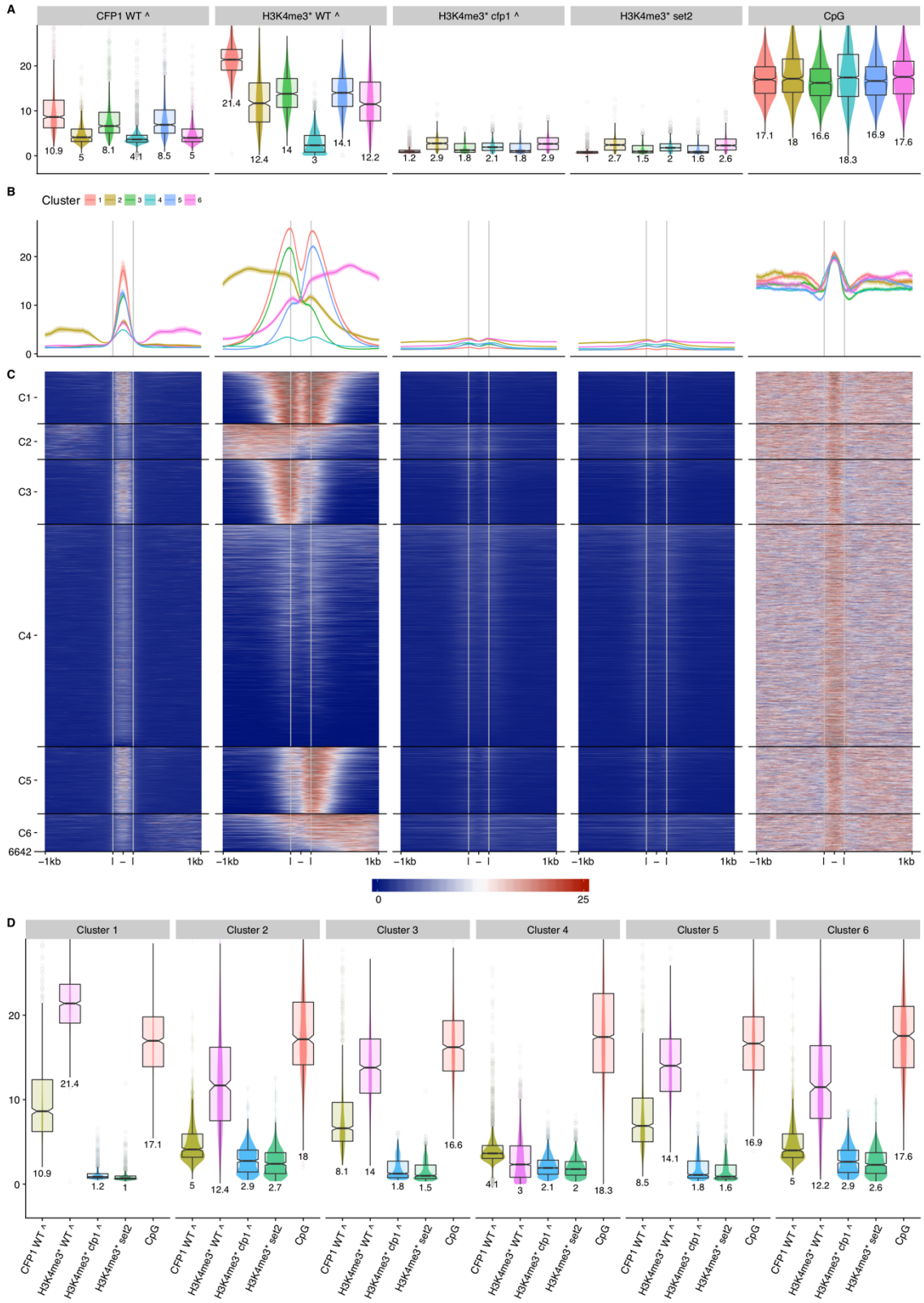


Figure 49 Combined plot showing the quantification of depletion of H3K4me3 and CpG dinucleotide density in 200bp window on CFP-1 peak sites in *cfp-1* and *set-2* mutant backgrounds. CFP-1 peak calls were divided into 6 clusters, showing diverse pattern of H3K4me3 marking. "^" denotes tracks used for clustering and "*" denotes tracks normalized with *C. briggsae* chromatin spike-in method. Panes A and D show p-values U-test. The numbers below box plots represent the mean of the quantified signal.

Cluster 5 is forward strand monodirectional promoter and exhibits opposite pattern of CpG density and chromatin marks - there is an asymmetric depletion of CpG density upstream CFP-1 binding site and strong H3K4me3 marking of +1 nucleosome, followed by asymmetric downstream marking of H3K36me3.

I observed that the directionality of H3K4me3 marking on +1 nucleosome is indeed well correlated with directionality of H3K36me3 marking. Further, there is a CpG depletion on -1 nucleosome, in opposite direction to H3K36me3 marking. I speculate that C1 cluster represents bi-directional promoters, while clusters C3 and C5 represents uni-directional expression. Cluster C4 does not seem to be active, and might represent regulated, or tissue specific promoters. Cluster C2 and C6 exhibit weak H3K4me3 marking, in line with my previous speculation that they represent alternative promoters and marking pattern indicates proximal, secondary promoter sites.
This observation allows me to speculate that nucleosomes marked with histone 3 lysine 4 trimetylation attract the polymerase and drive the expression directionality. Also, CpG depletion might be a barrier for polymerase, inhibiting productive elongation in this direction. Furthermore, in uni-directional promoters histone opposite to elongation direction might be marked with a different mark that further inhibits transcriptional elongation.

3.13 The CGC/GCG tri-nucleotides show strong enrichment and reverse complement imbalance at all promoters

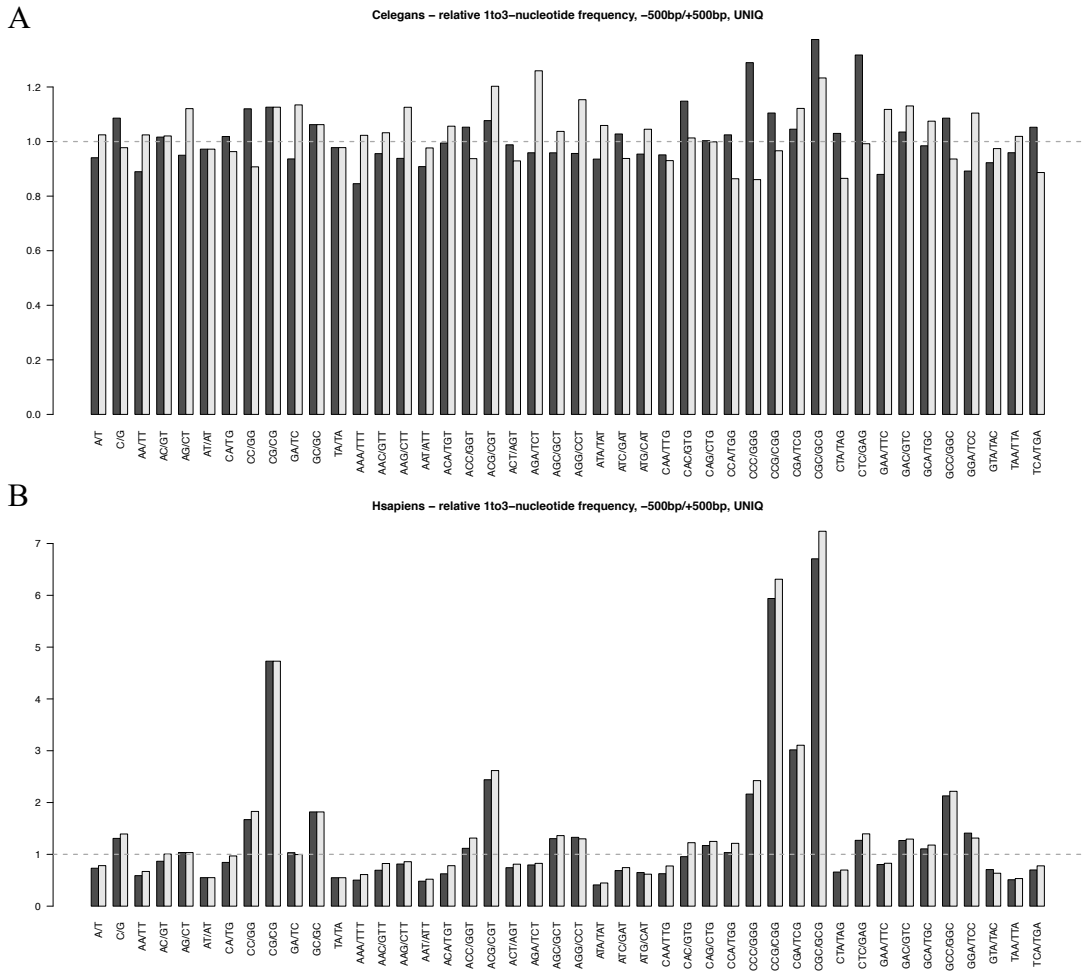


Figure 50 Bar plots illustrating sequence imbalance in promoters between k-mer (dark grey bar) and it's reverse complement sequence (light grey bar) for *C. elegans* (A) and *H. sapiens* (B). The values represent relative enrichment of k-mer in the promoter over genomic average. Promoter region is defined as non-overlapping regions +/-500bp from annotated TSS.

The above analyses showed that CpG density is important feature of HOT regions and that the CXXC1/CFP-1 protein binds CpG dense promoters. To investigate CFP-1 binding further, I asked if a trinucleotide motif might be even better indicator of HOT regions in *C. elegans* and *H. sapiens*. I found that CGC/GCG tri-nucleotides have a much stronger association with promoters in both organisms than CpG dinucleotides (**Figure 50**). Similarly, to CpG density analyses vs. GC content, I wanted to be sure that CGC/GCG tri-nucleotides density is not simply driven by CpGs. The CGC/GCG

observed vs. expected ratio, taking into consideration both GC and CpG content (see methods) supported the observation that CGC/GCG density is higher than expected.

In addition, both examined species shown the sequence imbalance between k-mer and its reverse complement sequence. In both species there is significant imbalance between CGC and GCG, but there are also other imbalances specific for the given species. **Figure 50** illustrates this phenomenon for *C. elegans* and human.

3.14 Differential expression analyses in *cfp-1* and *set-2* backgrounds

Initially it was observed, that "active genes are tri-methylated at K4 of histone H3" (Santos-Rosa *et al.* 2002). This observation led to common misconception, that H3K4me3 is a activating mark, hence proteins involved in its deposition should be transcriptional activators (Howe *et al.* 2017; Voo *et al.* 2000). Indeed, this misconception is so prevalent, that single line description of human CFP1/CXXC1 protein in UniProt database says, "Transcriptional activator that exhibits a unique DNA binding specificity for CpG unmethylated motifs with a preference for CpGG". However, there is no evidence of showing causality between SETDB1/COMPASS complex or H3K4me3 deposition and transcriptional activators. Moreover, recent studies show, that in ES cells the CFP1 loss of function and subsequent loss or alteration of H3K4me3 marks show both upregulation and downregulation of genes, with majority being up-regulated (Clouaire *et al.* 2014).

To investigate the link between COMPASS dependent H3K4me3 deposition and gene expression regulation, we profiled gene expression profiling in *cfp-1* and *set-2* mutant embryos. My colleague Yan Dong performed RNA-seq for three biological replicates for wild type and *cfp-1* and two replicates for *set-2* (**Table 12**). Because *C. elegans*

embryos cannot be tightly synchronized, she determined the range of embryo ages of the collections by DAPI staining.

ContactExpID	LibraryType	Strain	Stage
rYD076	polyA	N2	Emb
rYD077	polyA	N2	Emb
rYD078	polyA	cfp-1	Emb
rYD079	polyA	cfp-1	Emb
rYD102	polyA	cfp-1	Emb
rYD101	polyA	N2	Emb
rYD103	polyA	set-2	Emb
rYD104	polyA	set-2	Emb

Table 12 The list of embryo samples produced for *cfp-1* and *set-2* differential expression analyses.

To assess reproducibility of data, I performed principal component analyses (PCA) on protein coding gene expression estimates transformed with regularized log (Love *et al.* 2014). I first collected count data using HTseq and normalized them using regularized logarithm transformation (rlog). Regularized logarithm transforms count data to the log2 scale in a way that minimizes differences between samples for rows with small counts, and which normalizes with respect to library size (Love *et al.* 2014; Tsurumi & Li 2012). The rlog transformation produces a similar variance stabilizing effect as variance stabilizing transformation (VST) (Anders & Huber 2010), though rlog is more robust in the case when the size factors vary widely.

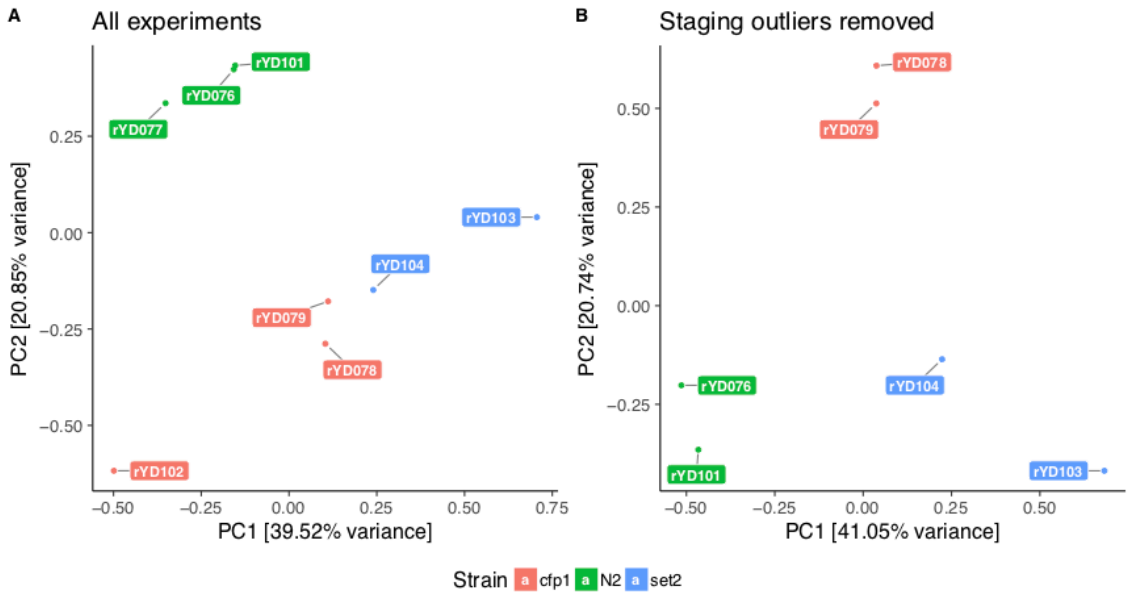


Figure 51 Principal component analyses (PCA) plot showing improvement of RNA-seq replicate matching after removing staging outliers. (A) PCA showing all RNA-seq experiments (B) improved PCA after removing outlaying rYD102 and rYD077 experiments. The PCA plot is based on rlog transformed read counts, done with HTseq method on STAR aligned data.

Upon initial inspection of expression data in the PCA plot (**Figure 51**A), I found that *cfp-1* sample rYD102 was an outlier, widely separated by principal components 1 and 2 (PC1 and PC2) from other *cfp-1* samples. Also, rYD077 was relatively separated from other two wild type samples. After inspecting the staging data (**Figure 52**) it became apparent that the outlying samples rYD102 and rYD077 have a larger proportion of older embryos than other samples collections (enriched for comma – 3-fold stage embryos). After removing these samples from analyses, the separation assessed by PCA plot was improved (**Figure 51**B). In further analyses I used the datasets without samples rYD102 and rYD077, leaving two replicates for each mutant.

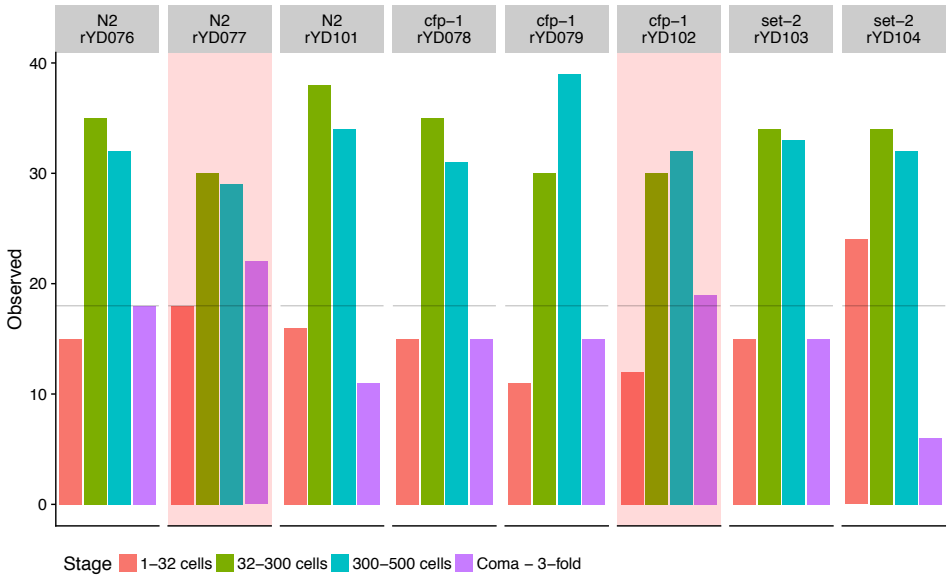


Figure 52 Staging of CFP-1 experiments presented as bar plot. The counting of worms in each stage was performed by Yan Dong. For each experiment 100 worms were selected and classified into following stages of embryo development – 1 – 32 cells, 23-300 cells, 300 – 500 cells, coma - 3-fold stage. This histogram shows the numbers of *C. elegans* observed in each state and each experiment. Outliers for WT and cfp-1 identified during PCA are older than other samples. The outlying samples are highlighted in red, the horizontal guide line shows the highest number of "coma - 3-fold stage" worms in the WT sample accepted for further RNA-seq analyses.

3.15 Significantly misregulated genes in *cfp-1* and *set-2* show little overlap with CFP-1 binding sites in promoter regions

Significant FDR>0.05	**Upregulated LFC>0**	***cfp-1* genes**	***set-2* genes**	**Overlaps**	**Percentage overlaps**
NS	down	9476	8792	6093	50.05%
NS	UP	10233	10497	7486	56.52%
YES	down	292	613	104	13.06%
YES	UP	174	273	62	16.31%

Table 13 Number of genes up- and down-regulated in *cfp-1* and *set-2* mutant backgrounds. "NS" mean non-significant (FDR>0.05). The table only summarises genes, where we can obtain an estimate for fold change and false discovery rate, so genes showing non-detectable expression are excluded.

To understand the connection between CFP-1 and its misregulated targets I have performed differential expression (DE) analyses on *cfp-1* and *set-2* RNA-seq data using DESeq2. I found 292 significantly downregulated and 174 significantly upregulated genes in *cfp-1* mutant embryos, and 613 significantly downregulated and 273 significantly upregulated in *set-2* mutant embryos (**Table 13**, false discovery rate (FDR)

< 0.05, log fold change (LFC) < or > 0 for downregulated and upregulated). I observed a significant overlap between genes upregulated and downregulated in *cfp-1* and *set-2* mutants – p-value equal $5.26*10^{-79}$ for up-regulated and $2.82*10^{-96}$ for down-regulated genes, as expected from their shared roles in H3K4me3 deposition and the membership in the COMPASS complex (**Figure 55**).

Interestingly, the numbers and the RNA-seq summary plots (correlation scatter plot, MAplot and Volcano plot presented for *cfp-1* in **Figure 53** and for *set-2* in **Figure 54**) show a bias for significantly downregulated genes in both mutant backgrounds. This contrasts with the bulk analyses of CFP-1 target genes (i.e. genes with promoters overlapping CFP-1 peaks), which are often upregulated in *cfp-1* mutant background.

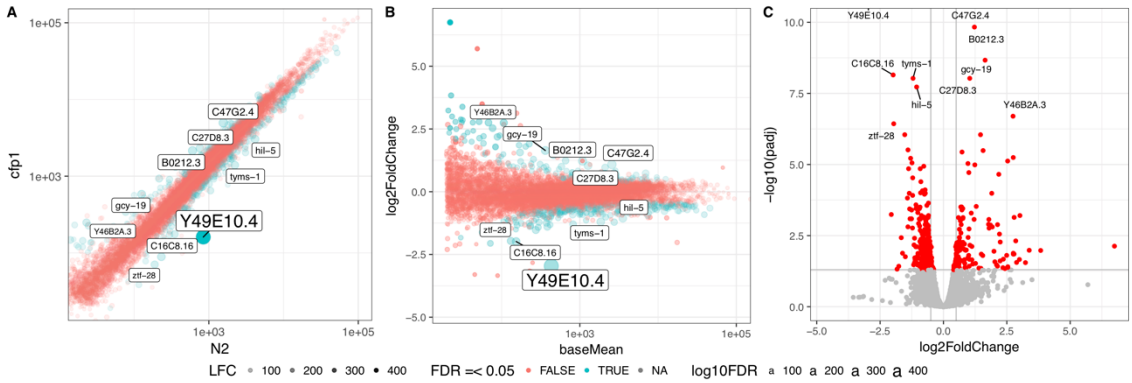


Figure 53 RNA-seq expression analyses for *cfp-1* mutant

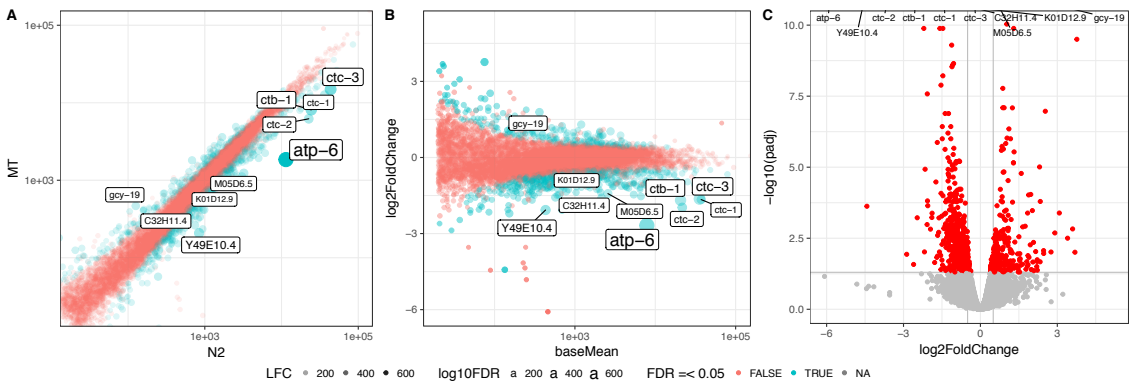


Figure 54 RNA-seq expression analyses for *set-2* mutant

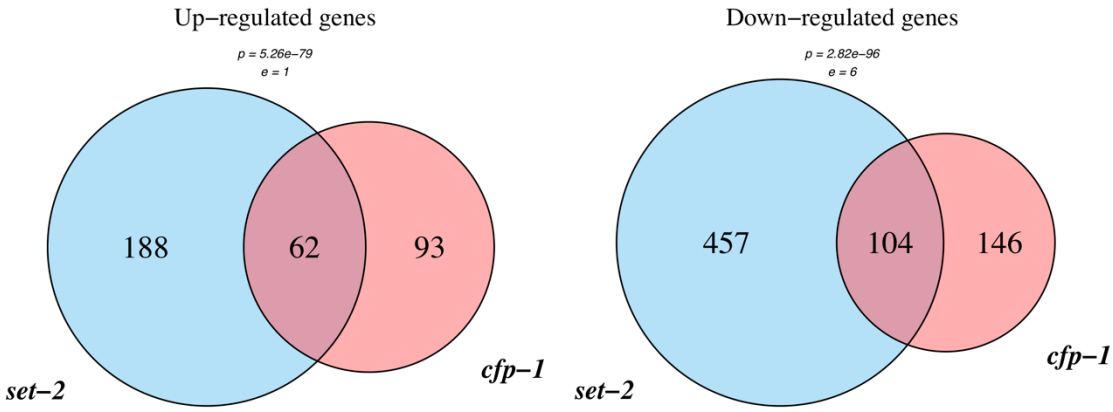


Figure 55 There is a small, but significant overlap between up- and down-regulated genes in *cfp-1* and *set-2* background. p-values (p) estimated using Fisher's test and expected values (e) are derived form hypergeometric distribution.

Next, I tested if there is an overlap between genes mis-regulated in *cfp-1* and *set-2* backgrounds, and if these genes are CFP-1 targets. Although it tested significant (Fisher's test p-value equal 5.26e-79 for up-regulated and 2.82e-96 for down-regulated genes, **Figure 55**), the total number of genes misregulated in both mutant backgrounds is low (**Table 13**), and most of mis-regulated genes do not overlap (only 62 out of 380 were up-regulated in any of two mutants, and 104 out of 796 were down-regulated in any of two mutants). The observation that most genes do not overlap suggest that despite the requirement for both genes for deposition of H3K4me3, they might also have distinct roles in gene regulation.

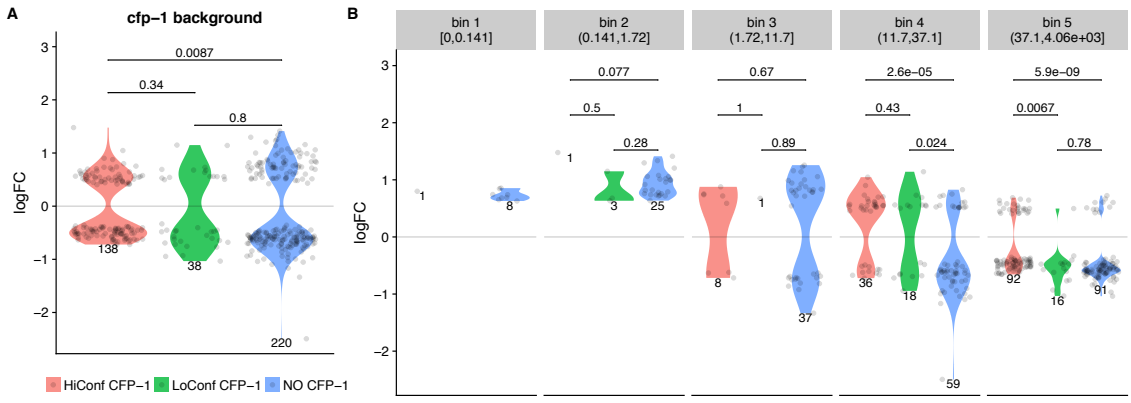


Figure 56 CFP-1 targets show weak association with both up- and down-regulated genes in *cfp-1*

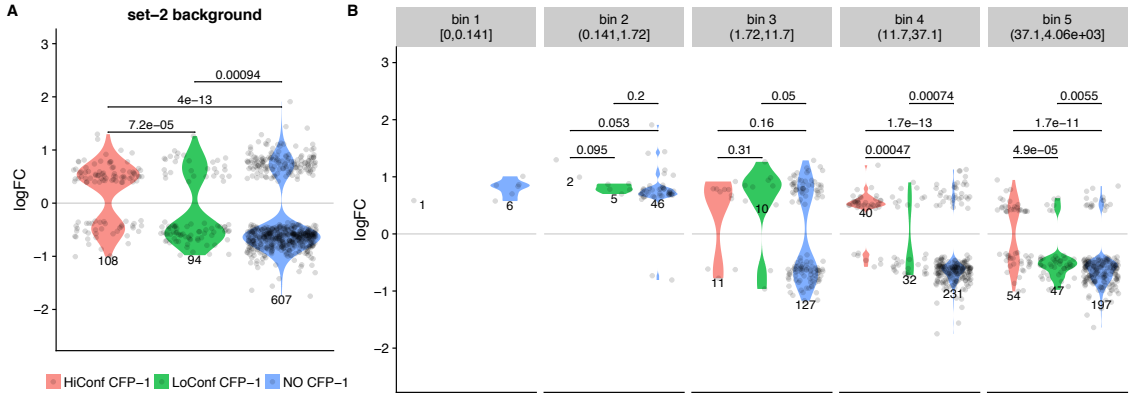


Figure 57 CFP-1 targets tend to be significantly up-regulated in comparison to non-target genes in *set-2* background.

There are many more high and low confidence CFP-1 targets than genes significantly up- or downregulated (466 versus 6642). I next investigated the relationship between CFP-1 binding and the individual genes found to be significantly misregulated in *cfp-1* or *set-2* mutants. For this analysis I have excluded downstream genes in operons (i.e. second and further genes in operon), as they are likely to be regulated from a single, upstream promoter and will be lacking CFP-1 peak despite the fact that upstream promoter might still overlap with it. I found 57 genes mis-regulated in *cfp-1* background to be downstream genes in operons. After removing them I had a set of 405 mis-regulated genes in *cfp-1* (**Table 14**). Then, I assigned promoters to all coding genes, as upstream 500 to downstream 500 base pairs, and selected promoters with HiConf and

LoConf CFP-1 peaks and not being annotated as downstream in operon – this gave me a set of 4,003 promoters.

Then I checked how many up and down-regulated genes are marked by CFP-1 in these promoter regions. I observed, that around one third (32.8% and 36.1%) of significantly down- and up-regulated genes in *cfp-1* have a CFP-1 marking at promoter regions (**Table 14**). This becomes even higher when we looked at all peaks - 42.0% and 45.8% for significantly down- and up-regulated genes in *cfp-1.* Much smaller portion of genes is marked by CFP-1 genes in *set-2* mutant background – only 18.2% (6.8% for HiConf peaks) and 39.2% (27.2% for HiConf peaks) of significantly down- and up-regulated genes, respectively.

Strain	Significant FDR>0.05	Upreg. LFC>0	# genes	CFP-1 targets	HiConf targets	% CFP-1	% HiConf
cfp-1	NS	down	5578	1441	896	25.8%	16.1%
cfp-1	NS	UP	4688	1817	1277	38.8%	27.2%
cfp-1	*	down	250	105	82	42.0%	32.8%
cfp-1	*	UP	155	71	56	45.8%	36.1%
set-2	NS	down	5379	1393	898	25.9%	16.7%
set-2	NS	UP	4541	1856	1319	40.9%	29.0%
set-2	*	down	561	102	38	18.2%	6.8%
set-2	*	UP	250	98	68	39.2%	27.2%

Table 14 Around a third of significantly mis-regulated genes in *cfp-1* and small fraction of *set-2* significantly mis-regulated genes show CFP-1 peak at promoter. The numbers in this table are after filtering downstream genes in the operons, hence the numbers of misregulated genes are smaller in comparison to table **Table 13**.

However, looking from perspective of all promoters marked by CFP-1 only a very small fraction is significantly up- or down-regulated in any mutant background. There are 2206 promoters marked by HiConf CFP-1 peaks and 3293 promoters marked by any CFP-1 peaks and only 211 and 330 unique significantly up- or down-regulated genes in any mutant marked by HiConf and all CFP-1 peaks respectively (out of all 1045 mis-regulated genes). This means we have detected a significant change of expression only

in 10% of genes marked by CFP-1. Also, in *cfp-1* background the CFP-1 peaks seem to be associated both with up- and down- regulated genes (**Figure 56**), while in *set-2* background CFP-1 peaks also significantly overlap promoters of upregulated genes (**Figure 57** and **Figure 58**B), however the number of overlapping genes in very small, which might render this trend insignificant.

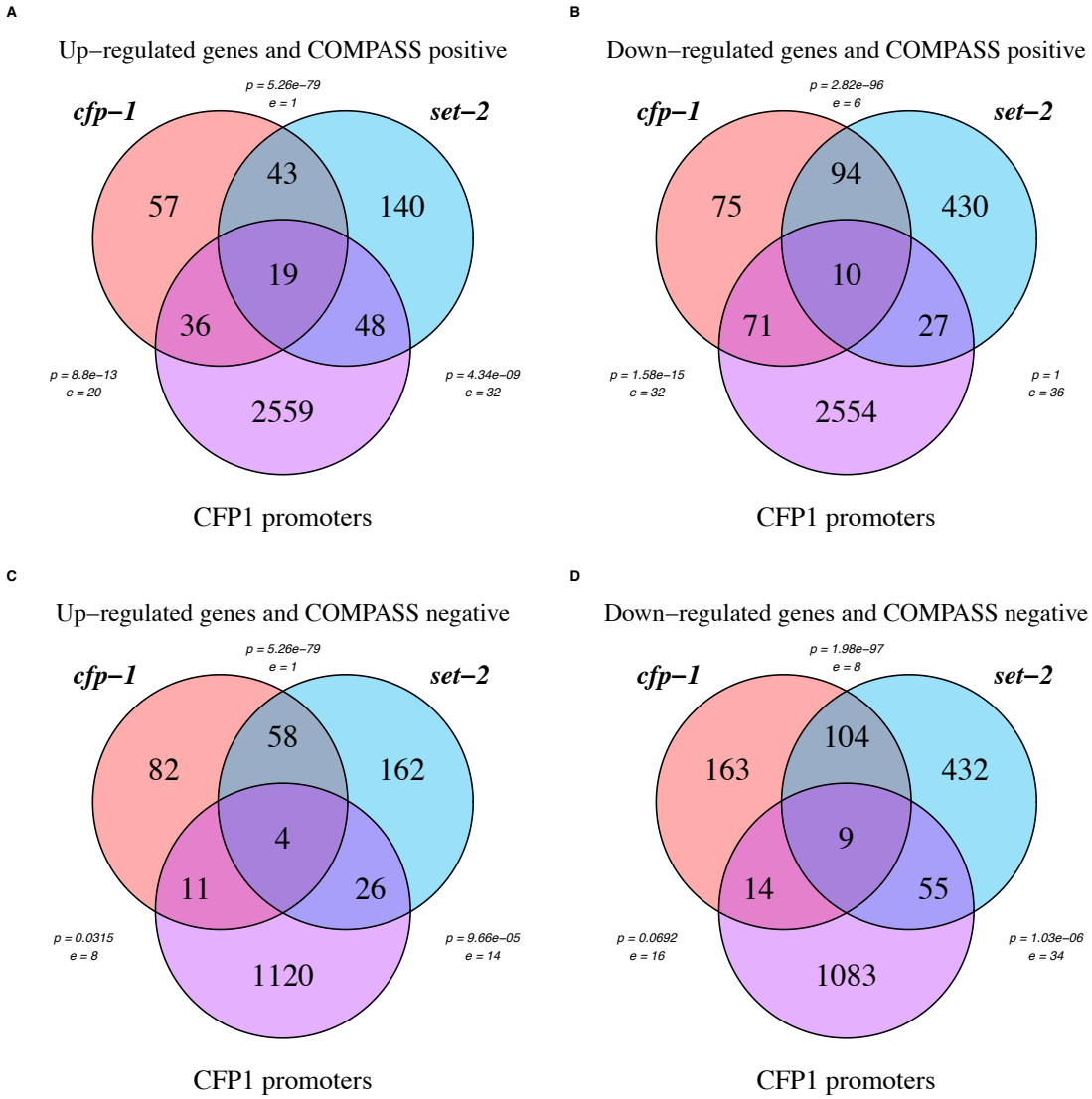


Figure 58 Genes mis-regulated in *cfp-1* and *set-2* mutant significantly overlap each other and weakly overlap CFP-1 targets under some circumstances. p-values (p) estimated using Fisher's test, and expected values (e) are derived from hypergeometric distribution. "COMPASS positive" promoters mean that these promoters overlap HiConf *cfp-1* peaks, while "COMPASS negative" denotes promoters overlapping low confidence CFP-1 peaks.

Further, I have tested if the small overlaps between CFP-1 marked promoters and differentially expressed genes in *cfp-1* and *set-2* backgrounds were significant. Given

the background of all coding genes (20,362) the most significantly overrepresented overlap was between genes down-regulated in *cfp-1* mutant and genes whose promoter regions are marked by CFP-1 (**Figure 58**). The overlaps between genes up-regulated in *cfp-1* and *set-2* background were less significant, and down-regulated in *set-2* background was not significant.

Given the background of coding genes with detectable expression (read count base mean > 100, 11,935 genes, **Figure 59**), which might be more appropriate background to accurately estimate p-value from Fisher test, only overlaps between genes mis-regulated in *cfp-1* and CFP-1 marked promoters tested as significant. This observation suggests that absence of CFP-1 shows weak association with up- and down-regulation of direct target genes, that is specific to *cfp-1* mutant, and not directly caused by loss of H3K4me3 methylation - lack of significant overlap/lower significance of overlaps in *set-2* mutant.

In conclusion, genes associated with the vast majority of CFP-1 sites did not individually show significant expression changes in either mutant background, but in bulk I found a small but significant increase in expression. Also, majority of differentially expressed genes are not marked by CFP-1, suggesting they are not direct targets. Both direct and indirect targets can be up- and down-regulated, suggesting that CFP-1 have some role in gene expression, but is not a transcriptional activator. Extrapolating and integrating these results with genes differentially expressed in *set-2* mutant, H3K4me3 is not an activating mark, as its targets can go up or down in terms of gene expression.

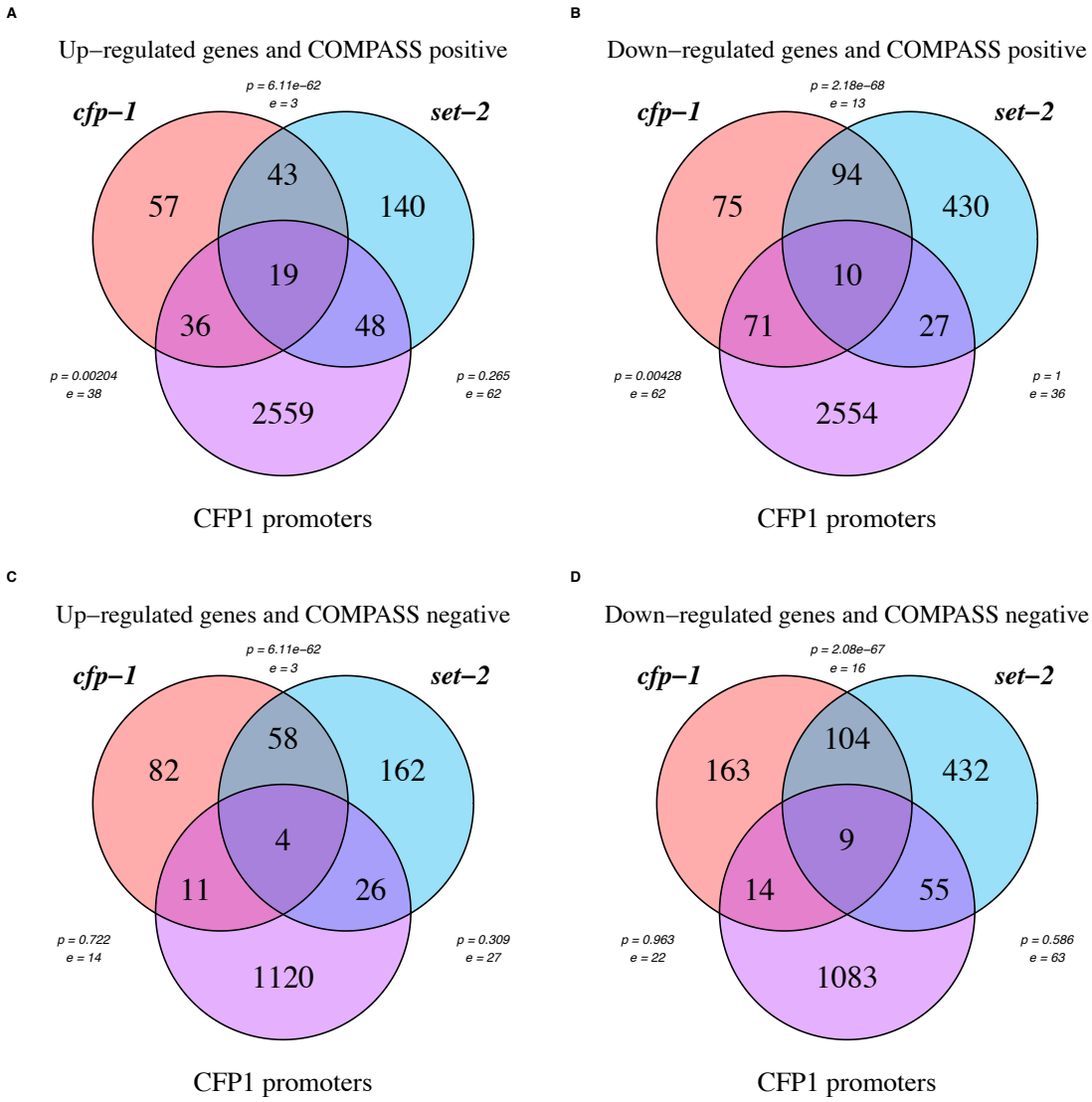


Figure 59 Overlaps between genes mis-regulated in *cfp-1* and CFP-1 marked promoters tested as significant in background of coding genes with detectable expression (read count base mean > 100, 11,935 genes). p-values (p) estimated using Fisher's test and expected values (e) are derived from hypergeometric distribution. "COMPASS positive" promoters mean that these promoters overlap HiConf *cfp-1* peaks, while "COMPASS negative" denotes promoters overlapping low confidence CFP-1 peaks.

This observation is consistent with previous reports in mouse. Murine ES cells mutant for Cfp1 (mouse ortholog gene) also displayed few gene expression changes and no clear association between binding and expression changes (Brown *et al.* 2017a; Clouaire *et al.* 2014).

3.16 CFP-1 targets are up-regulated in *cfp-1* and *set-2* backgrounds in comparison to genes not marked by CFP-1 in promoter regions

Further, I investigated the relationship between CFP-1 binding and gene expression changes. For this I have calculated shrunken log fold change (Love *et al.* 2014) between mutant backgrounds and wild type samples. I divided coding genes into three categories based on the association with a CFP-1 peak in the promoter region (see Methods): (1) those with a high confidence/COMPASS CFP-1 peak (HiConf, 2729 genes), (2) those with a low confidence CFP-1 peak (LoConf, 1258 genes) and (3) those without a CFP-1 peak in the promoter region (13759 genes). I used the CFP-1 peak sets and clusters I previously created (**Figure 43**).

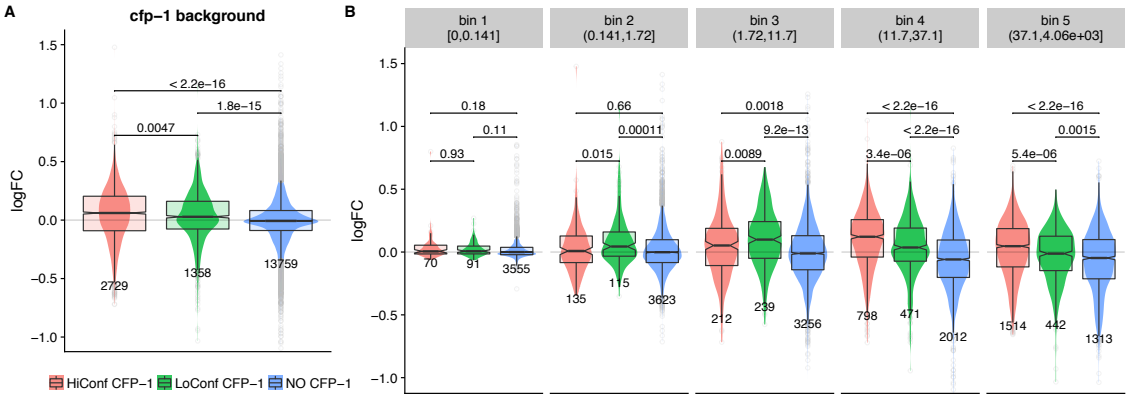


Figure 60 CFP-1 targets are up-regulated in *cfp-1* background in comparison to genes not marked by CFP-1 in promoter regions.

This analysis shows that the groups of HiConf and LoConf CFP-1 targets show a small but significant upregulation in both in *cfp-1* and *set-2* mutant backgrounds (p-value < 10e-10 by Mann-Whitney U-test) in comparison to targets not associated with CFP-1 (**Figure 60**A and **Figure 61**A). HiConf targets are also significantly upregulated in comparison to LoConf targets (p-value of 0.0047 for *cfp-1* and 2.5e-6 for *set-2* by Mann-Whitney U-test). To make sure that the effects were not caused by small upregulation of lowly expressed genes, I divided the expression data into five expression bins based on wild type expression – fragments per kilobase of transcript per

million transcript in library (FPKM) for these bins are shown in **Figure 60**B and **Figure 61**B.

I observed that upregulation in both Hiconf and LoConf sets is driven changes in genes in the higher quintiles of expression, with fourth quintile showing the strongest effect for HiConf targets. Interestingly, the middle quintile of expression (1.72 to 11.7 FPKM) shows the strongest upregulation of LoConf genes, and only weak for HiConf genes – with an opposite trend in comparison to two top quintiles of expression. Genes in the two lowest quintiles of expression (FPKM =< 1.72) show no significant change of CFP-1 targets in comparison to other genes (**Figure 60**B and **Figure 61**B). These results suggest that CFP-1 and SET-2 have a repressive effect on gene expression.

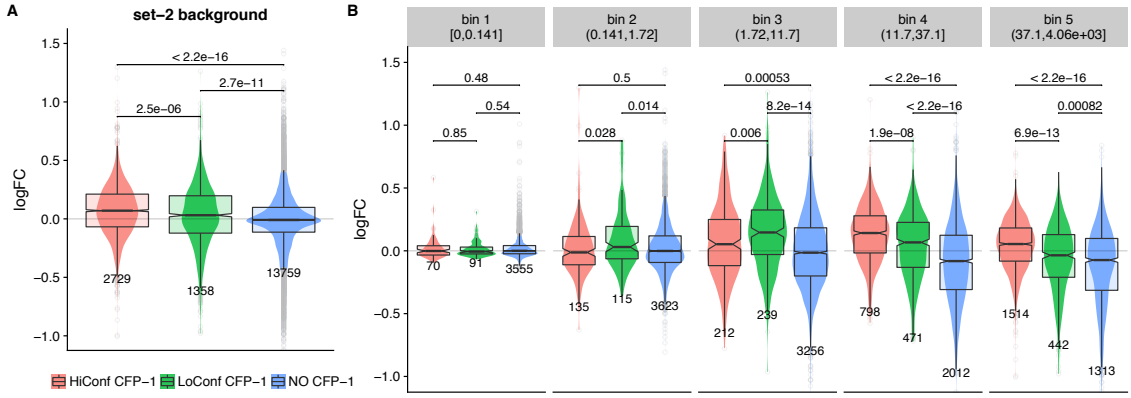


Figure 61 CFP-1 targets are up-regulated in *set-2* background in comparison to genes not marked by CFP-1 in promoter regions

3.17 CFP-1 driven H3K4me3 deposition might have a role in stabilizing gene expression

Previous analyses have shown that *cfp-1* and *set-2* mutants show both up-and downregulation of coding genes, with a slight bias for upregulation. In the next step I wanted to test if CFP-1 dependent H3K4me3 deposition might have a stabilizing role on

gene expression. Stabilizing role of H3K4me3 has been also proposed in literature (Clouaire *et al.* 2014; Howe *et al.* 2017).

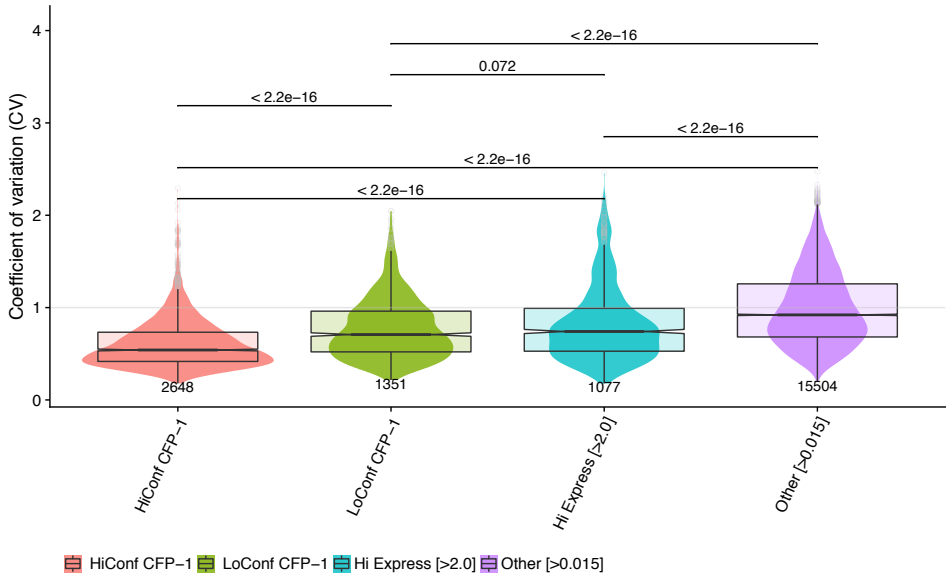


Figure 62 HiConf CFP-1/COMPASS target genes are more stably expressed during embryo development than non CFP-1 target genes. p-values shown above boxplots represent statistical significance calculated using Mann-Whitney U-test.

To explore this hypothesis, I investigated whether genes associated with CFP-1 binding have broad stable expression versus being developmentally regulated. To do this, I calculated the coefficient of variation (CV) of gene expression using an embryo developmental time course RNA-seq dataset (Boeck *et al.* 2016). Genes that are broadly and stably expressed across embryogenesis have low CV values and genes that are expressed at particular times in embryogenesis have high CV values. I observed that HiConf CFP-1/COMPASS targets have significantly lower CV values in comparison to highly expressed genes or all genes, showing that their expression is more stable during embryo development (**Figure 62**). Also, the LoConf targets have lower CV values than non-CFP target genes or all highly expressed genes, but they have higher CV values than HiConf CFP-1/COMPASS targets (**Figure 62**).

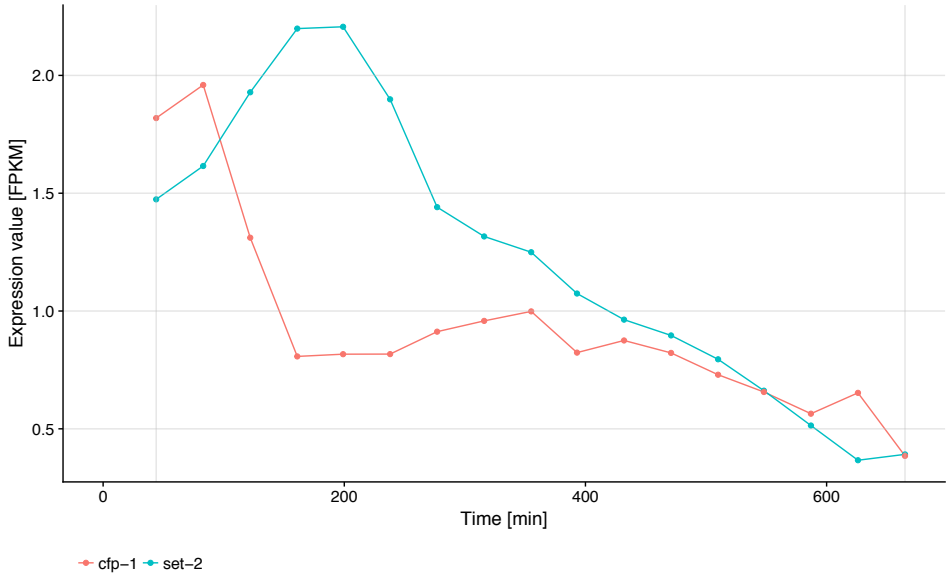


Figure 63 Expression values (FPKMs) of CFP-1 and SET-2 during embryo development (44min - 665min).

To further investigate what happens to these genes I plotted fragments per kilobase of transcript per million reads in the library (RPKM) values of expression as the function of time. The examination of CV values showed that analysed classes show different stability of time dependant expression profiles. RPKM analyses give me information if time dependant expression goes up, down or show other type of fluctuations. To gain intuition when CFP-1 and SET-2 is expressed I started with plotting expression values of just these two genes as the function of time. Both SET-2 and CFP-1 are expressed early in embryo development, with peak expression for CFP-1 at 83min and peak expression for SET-2 between 161 and 199min (**Figure 63**).

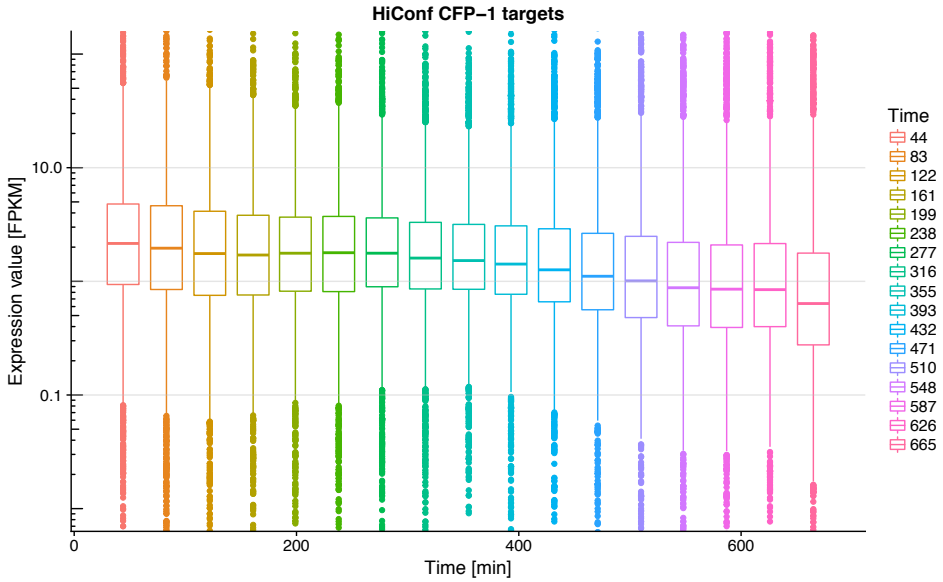


Figure 64 HiConf/COMPASS CFP-1 peaks associated genes start with high expression and generally stays expressed with slight bias for genes going down with the expression.

Further, I have divided coding genes in same classes as presented at **Figure 62**, and plotted their expression summary statistics as function of time using a series of boxplots (**Figure 63**, **Figure 64**, **Figure 65** and **Figure 66**). I observed genes associated with HiConf/COMPASS CFP-1 peaks are high expressed and generally remains highly expressed throughout embryo development with a slight bias for genes being down-regulated in late embryo stages (**Figure 64**). Genes marked with LoConf CFP-1 peaks at promoter regions are lowly expressed in early embryo development and tend to increase their expression as development progresses (**Figure 65**). Highly expressed genes tend to stay at high expression level, showing higher variance in early and late embryo development stages (**Figure 66**). Finally, lowly expressed genes tend to gain expression in late embryo stages (**Figure 67**). Taken together, this might suggest that HiConf and LoConf associated genes might be regulated differently during development.

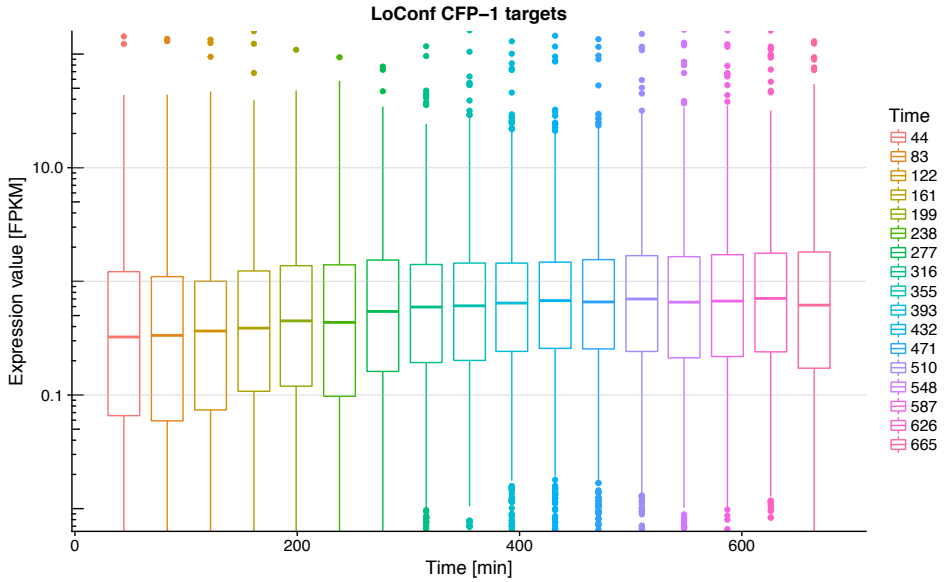


Figure 65 LoConf CFP-1 associated genes show lower expression than HiConf CFP-1 genes and tend to go up during embryo development.

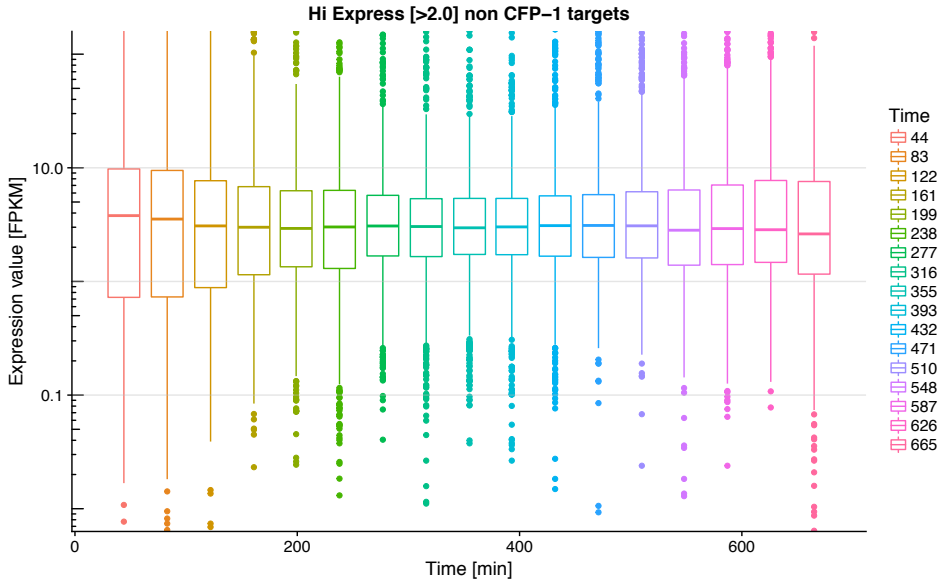


Figure 66 Highly expressed genes not associated with CFP-1 tend to stay at the same level of expression level, with high variance at early and late development.

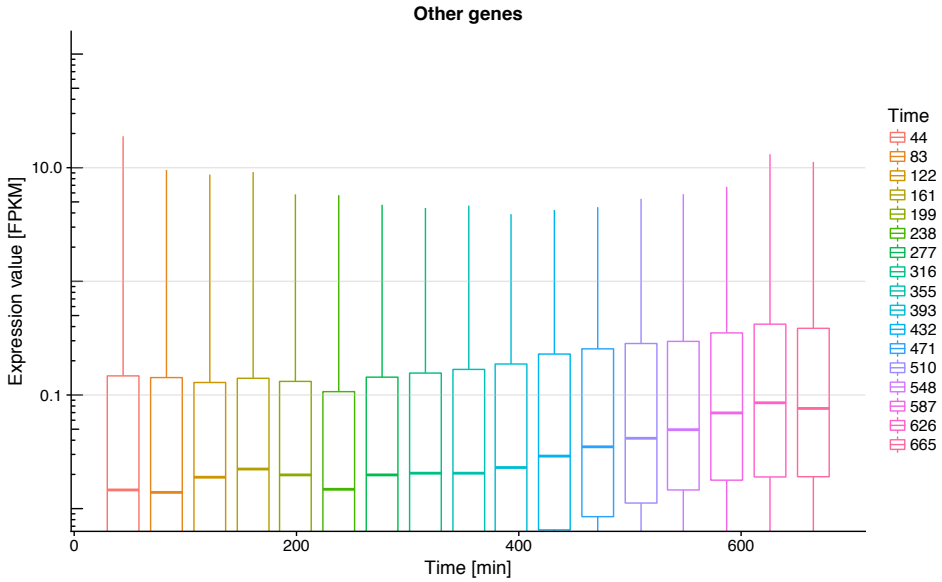


Figure 67 Lowly expressed genes not associated with CFP-1 tend to show relatively high increase in expression in the late embryo development.

Further, I was interested if significantly, differentially regulated genes in both mutant backgrounds show any specific developmental profile of regulation. To answer this question, I first plotted CV values of up- and down- regulated genes in cfp-1 and set-2 mutant backgrounds and compared them to highly expressed genes and all genes. I observed, in both mutants, up-regulated CVs are no different than highly expressed genes and slightly smaller than lowly expressed ones (**Figure 68**). Contrary, the CV values in down-regulated genes are much higher than any other class, which suggests they are developmentally regulated in embryo (**Figure 68**). Also, the CV of upregulated genes in *set-2* mutant is much higher than in *cfp-1*.

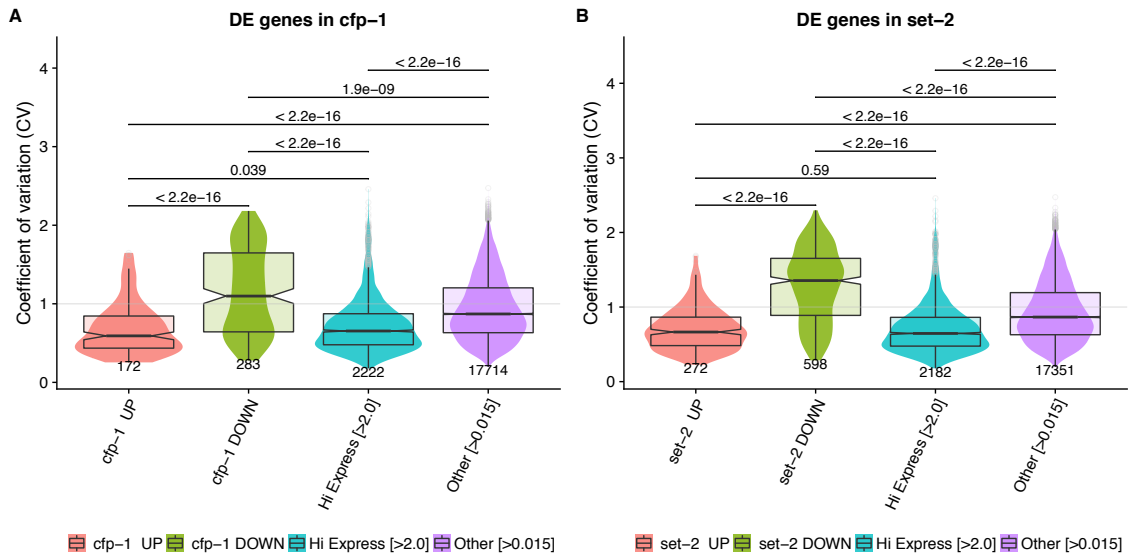


Figure 68 In *cfp-1* and *set-2* mutants up-regulated genes show similar CV to highly expressed ones, but down-regulated genes expression is significantly more variable than in any other class.

Then, similarly to CFP-1 targets vs. other genes I wanted to ask if differentially regulated genes show any special profile of expression in development. I observed in both mutants up-regulated genes show relatively stable expression profile throughout embryo development, with a slight bias for higher expression in middle embryo development (**Figure 69** and **Figure 70**), which is consistent with CV values (**Figure 68**). Contrary, down-regulated genes show different profiles of embryonic expression. In *cfp-1* mutant down-regulated genes show higher expression toward beginning and end of embryo development, and lower during middle stages (**Figure 69**). In *set-2* mutant down-regulated genes start with low expression and exhibit very significant and stable growth of expression in later embryo stages.

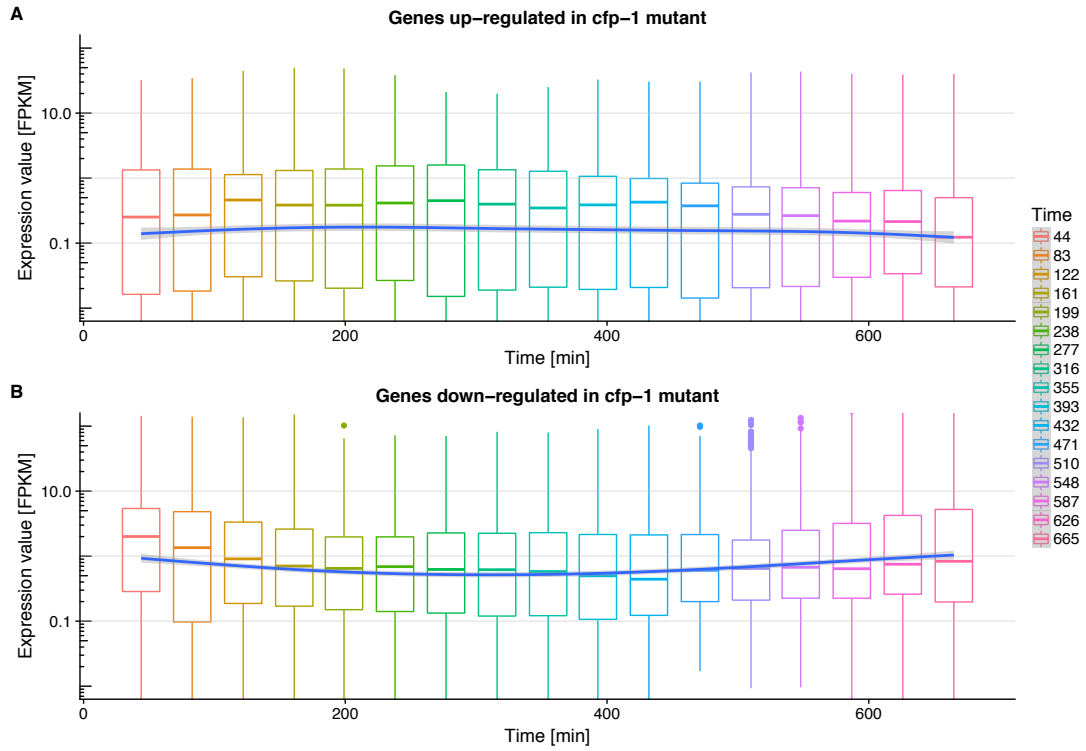


Figure 69 The embryo developmental profile of up- (A) and down-regulated (B) genes in *cfp-1* mutant background. Upregulated genes show relatively stable profile of expression thought embryo development, while down-regulated genes show higher expression towards beginning and end of embryo development. Blue line represents generalized additive models with integrated smoothness estimation – a method to visualise smoothed, robust estimate of continuous expression based on sampled interval data points. Semi-transparent filed represents estimation error.

Taken together this data shows that CFP-1 and SET-2 dependant H3K4me3 deposition might play the role in stabilising expression during development, as genes with promoters marked by CFP-1 tend to be more stably expressed. Also, a trend of down-regulated genes in *set-2* mutant increasing expression through development is quite interesting and might suggest that we are observing them as down-regulated, because their failed to increase their expression to wild-type levels because lacking SET-2 function. However, I cannot draw strong conclusions without having a time resolved transcriptome in these mutant backgrounds. Hence, further experiments are required to fully understand the role of CFP-1 and SET-2 and H3K4me3 histone modification in development.

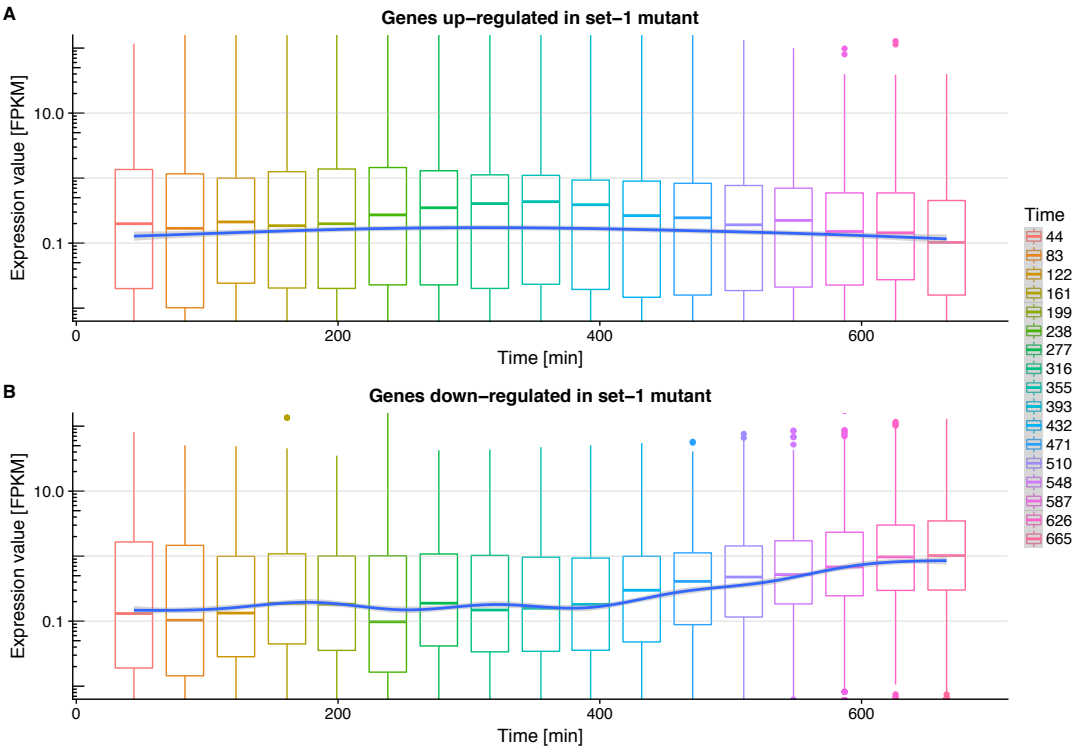


Figure 70 The embryo developmental profile of up- (A) and down-regulated (B) genes in *set-2* mutant background. Upregulated genes show a relatively stable profile of expression throughout embryo development, while down-regulated genes show a very significant and stable growth of expression in later embryo stages. Grey line represents generalized additive models with integrated smoothness estimation – a method to visualise smoothed, robust estimate of continuous expression based on sampled interval data points. Semi-transparent ribbon represents estimation error.

3.18 CFP-1 functionally interacts with the Sin3S/HDAC complex

The association of CFP-1 and SET-2 with both activation and repression of coding genes might be due to interaction with other chromatin modifying complexes. H3K4me3 has been reported to interact with different histone acetyltransferase (HAT) and histone deacetylase (HDAC) complexes such as Sgf29 of the histone acetyltransferase complexes SAGA (Bian *et al.* 2011; Vermeulen *et al.* 2010), BPTF in the NuRF chromatin remodelling complex (Li *et al.* 2006; Pinskaya *et al.* 2009; Santos-Rosa *et al.* 2003; Wysocka *et al.* 2006), the ING family proteins associated with either HAT or HDAC complexes (Lee *et al.* 2009; Martin *et al.* 2006; Shi *et al.* 2006; Taverna *et al.* 2006), and Tip60 of the HAT and nucleosome exchange complex (Taverna *et al.* 2006). In addition, human cell-proliferation factor HCF1 was shown to tether the

SETD1A and Sin3 histone deacetylase complexes together (Wysocka *et al.* 2003). Our collaborator Francesca Palladino and members of her laboratory recently found through proteomics analyses that *C. elegans* CFP-1 physically associates with components of the Sin3S HDAC complex. They also found that the *sin-3* mutant phenotype resembles that of cfp-1 mutants and that the two genes interact. It should be noted that SIN-3, similarly to SET-2 and CFP-1, plays a role in deposition of H3K4 methylation. However, in contrast to MET-2, SIN-3 is specific for H3K9me2 deposition on asynapsed chromosomes (Checchi & Engebrecht 2011). In the next section, I describe experiments investigating the interaction between CFP-1 and SIN-3.

The Palladino lab co-purified four components of the Sin3S complex with CFP-1: SIN-3, ATHP-1, HDA-1/HDAC1 and MRG-1. Our lab mapped the binding profiles of SIN-3 and HDA-1 in wild-type embryos to compare to CFP-1, and in *cfp-1* mutants to ask if CFP-1 played a role in their recruitment. I analysed the ChIP-seq experiments.

Factor	Strain	Stage	Condition	ExtractID	SampleID
HDA1	cfp-1	MxE	1	aa148	HDA1_cfp1_AA779_aa148
HDA1	cfp-1	MxE	1	aa150	HDA1_cfp1_AA781_aa150
SIN3	cfp-1	MxE	1	aa148	SIN3_cfp1_AA824_aa148
SIN3	cfp-1	MxE	1	aa150	SIN3_cfp1_AA825_aa150
HDA1	set-2	MxE	1	aa163	HDA1_set2_AA814_aa163
HDA1	set-2	MxE	1	aa164	HDA1_set2_AA815_aa164
SIN3	set-2	MxE	1	aa163	SIN3_set2_AA826_aa163
SIN3	set-2	MxE	1	aa164	SIN3_set2_AA827_aa164
HDA1	sin-3	MxE	2	aa159	HDA1_sin3_AA816_aa159
HDA1	sin-3	MxE	2	aa161	HDA1_sin3_AA817_aa161
HDA1	N2	MxE	1	aa147	HDA1_N2_AA778_aa147
HDA1	N2	MxE	1	aa149	HDA1_N2_AA780_aa149
HDA1	N2	MxE	2	aa160	HDA1_N2_AA812_aa160
HDA1	N2	MxE	2	aa162	HDA1_N2_AA813_aa162
SIN3	N2	MxE	1	aa147	SIN3_N2_AA761_aa147
SIN3	N2	MxE	1	aa149	SIN3_N2_AA762_aa149

Table 15 The summary of ChIP-seq experiments analysed for HDA-1 and SIN-3 functional analyses.

To investigate the relationship between CFP-1, SIN-3 and HDA-1 at a genomic level I compared their ChIP-seq binding profiles in wild type and determined whether HDA-1 or SIN-3 profiles were affected in *cfp-1*mutant embryos. **Table 15** lists the datasets that were generated by Alex Appert and Yan Dong in our lab.

To begin this study, I have analysed the pattern of HDA-1, SIN-3 and HCF-1 in context of previously analysed CFP-1 and H3K4me3 data. I called peaks separately on each replicate, and then combined them using interval intersection, producing a single peak call set per factor (see Methods). Then I calculated overlapping regions between CFP-1, HDA-1 and SIN-3. I also investigated associations between all five tested factors using interval union (similarly to "Any5" set discussed in Chapter 2), annotated them with factors immunoprecipitated in these regions, and converted this annotation to a Venn diagram (**Figure 71**B).

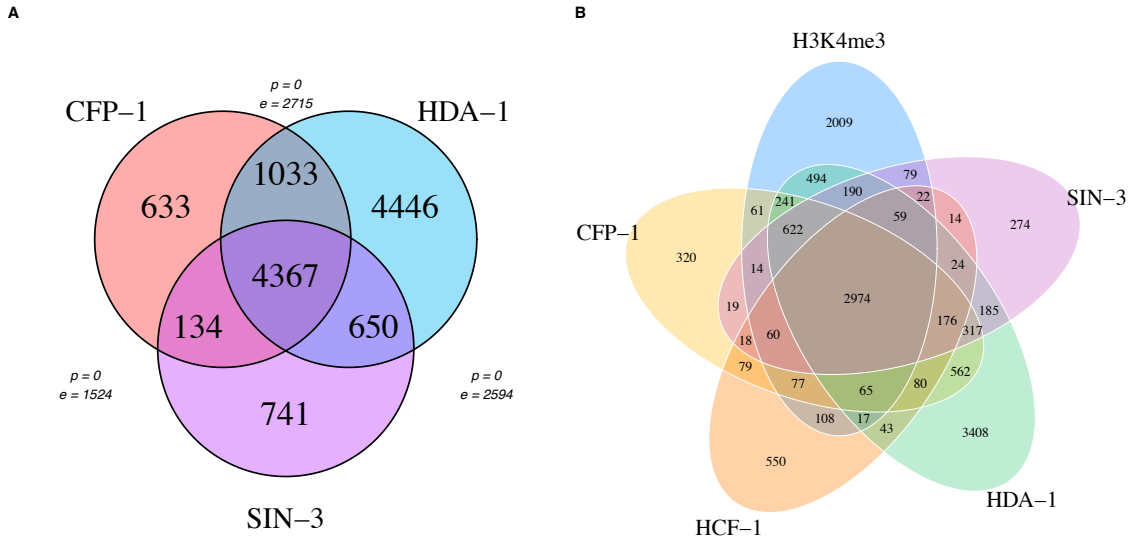


Figure 71 CFP-1, HDA-1 SIN-3, HCF-1 and H3K4me3 co-localise at many regions in the genome. (A) Three-way Venn diagram for CFP-1, HDA-1 and SIN-3, shows that majority of CFP-1 and SIN-3 sites overlap, and these sites usually also contain HDA-1. However, HDA-1 also binds to other locations (B) Five-way Venn diagram showing overlap between peak calls. In addition to HDA-1, SIN-3, CFP-1 and H3K4me4, HCF-1 (ortholog of human host cell factor) is also shown.

First, I observed that CFP-1, SIN-3, and HDA-1 show good overlap, with 4367 out of

12004 union sites (36%) occupied by all factors (**Figure 71**A). The majority of CFP-1 and SIN-3 sites overlap, and these sites usually also contain HDA-1. However, HDA-1 also has a considerable number of unique binding loci (4446; Figure 60A). Adding HCF-1 and H3K4me3 peaks to these analyses, I observed a large overlap between all five factors – 3971 out of 13161 sites (30%) are common. CFP-1, SIN-3 and HCF-1 tend to overlap with one or more other factors, while HDA-1 and H3K4me3 have a high number of unique sites – 3408 and 2009 unique sites respectively (**Figure 71**B). The high number of unique peaks for H3K4me3 might be detected because it exhibits a pattern of two peaks flanking the nucleosome depleted region (NDR), hence it might not overlap directly with factors binding in the middle of given NDR. The NDR centric (i.e. NDR plus flanking region) analyses would be needed to elucidate on this further.

3.19 SIN-3 abundance is reduced on high confidence CFP-1 regions in *cfp-1* mutant strain

I next analysed the signal on CFP-1 binding sites using the previous subdivision of high (C1, HiConf) and low CFP-1 occupancy (C2 - LoConf) peaks. Heatmap analyses confirmed the good overlap between CFP-1 and SIN-3 (**Figure 72**). Further, *cfp-1* mutants showed a clear reduction of SIN-3 levels in HiConf (C1) cluster whereas SIN-3 signal appeared unchanged in the LoConf cluster (C2). This was confirmed by signal quantifications (**Figure 72**). Finally, I have tested whether the loss of SIN-3 was statistically significant – indeed the major loss of SIN-3 signal in C1 tested very significant (**Figure 73**). Interestingly, C2 showed a very small gain in CFP-1 signal that also tested significant. I conclude, that in bulk analyses SIN-3 is reduced in HiConf CFP-1 sites in *cfp-1* mutant background.

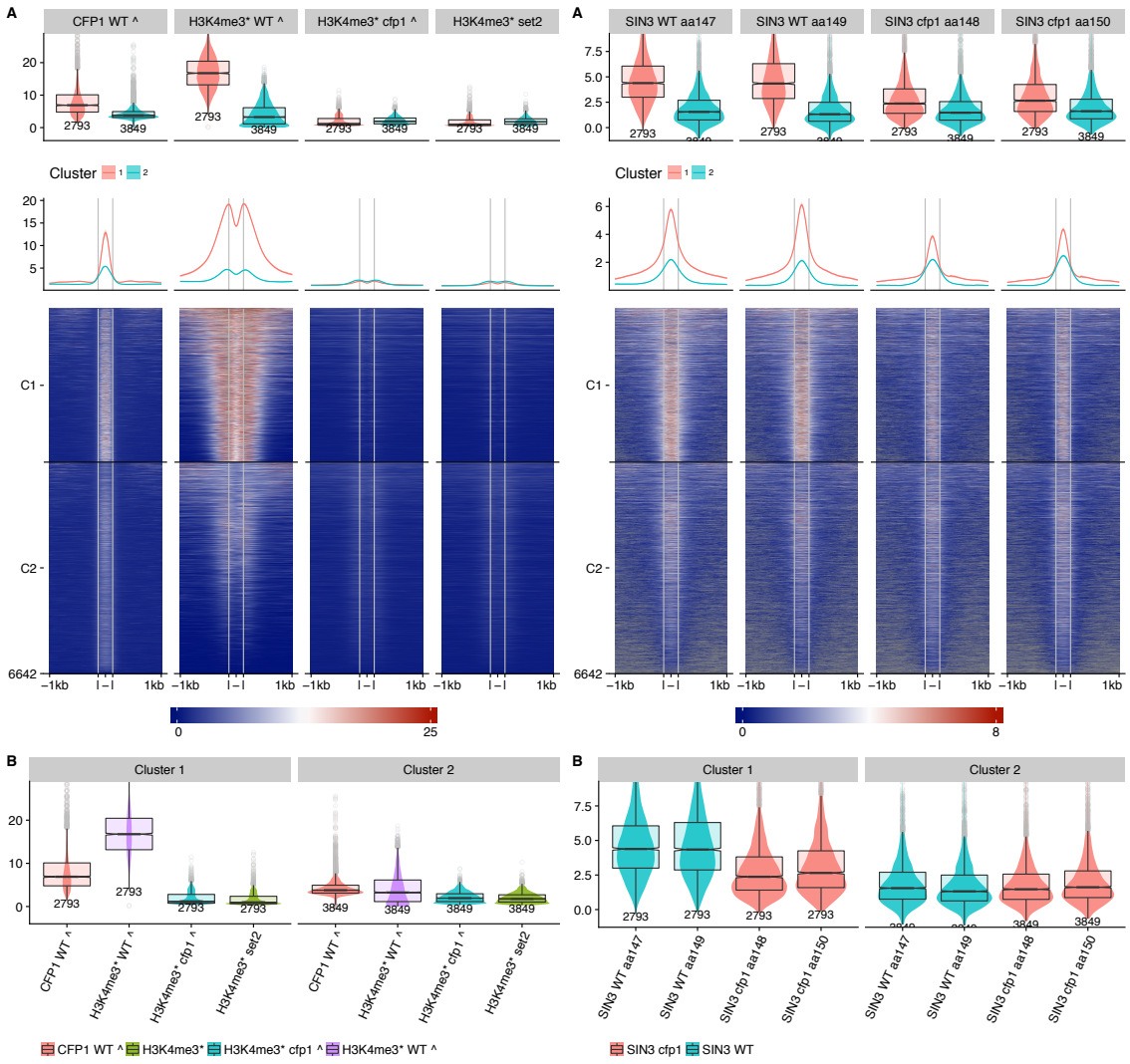


Figure 72 SIN-3 in ChIP-seq in WT and *cfp-1* background shown in CFP-1 HiConf (C1) and LoConf (C2) peaks in context of CFP-1 signal and H3K3me3 in WT, cfp-1 and set-2 backgrounds. The combined figure of quantifications, heatmaps and profile plots. Panel A-left shows boxplot quantifications, profile plot quantifications and heatmaps for CFP-1 and H3K4me3 in WT, *cfp-1* and *set-2* mutant backgrounds. Panel A-right shows same plots for SIN-3 in WT and *cfp-1* strains. Panels B-left and B-right show same quantifications as boxplots on panel A but are arranged in the way that makes it easy to compare replicates rather than clusters.

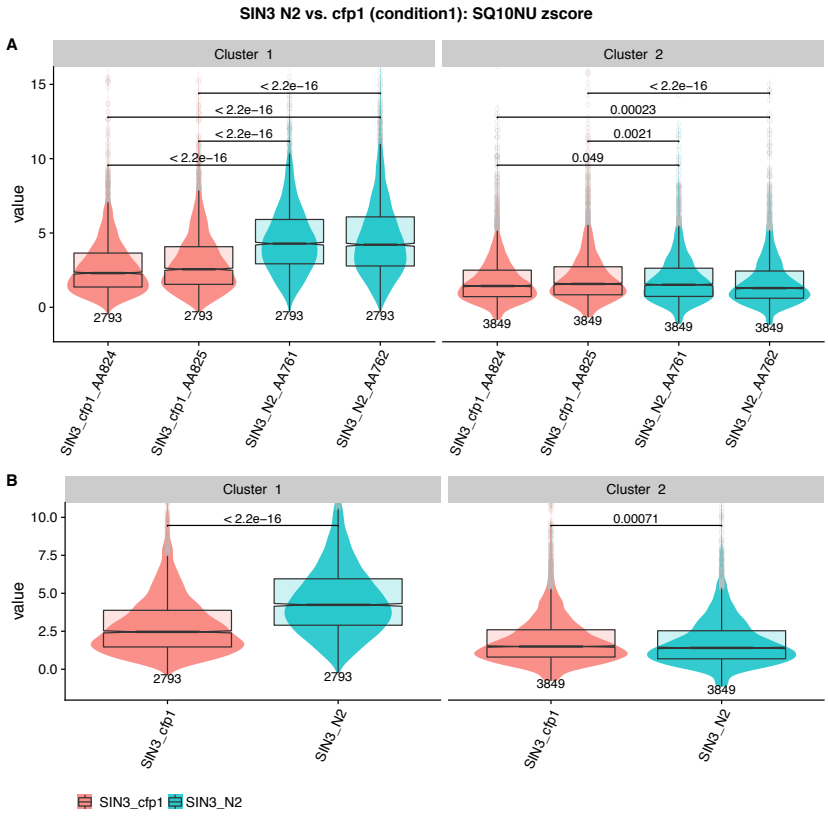


Figure 73 Significance testing for difference between SIN-3 in WT and *cfp-1* backgrounds. P-values are estimated using Mann-Whitney U test – a nonparametric method, which tests whether there is a location shift between two distributions (Bauer 1972).

3.20 HDA-1 abundance is reduced on high confidence CFP-1 regions in *cfp-1* mutant strain

I next asked whether the association of HDA-1 to CFP-1 sites depends on CFP-1, again analysing HiConf and LoConf CFP-1 bound regions separately. As observed for SIN-3, heatmap analyses confirmed the good overlap between CFP-1 and HDA-1 (Figure 68). Also as observed for SIN-3, I found that HDA-1 levels were reduced in the HiConf (C1) cluster which was confirmed by signal quantifications (**Figure 74**) and statistical testing (**Figure 75**). In contrast to SIN-3, there was a loss of HDA-1 signal in the LoConf cluster C2 in *cfp-1* mutants in comparison to WT that tested significant in combined replicates (**Figure 75B**). However, while testing replicates against each other (**Figure 75**A) it became evident that they did not match well in C2, so I consider this effect insignificant and non-conclusive.

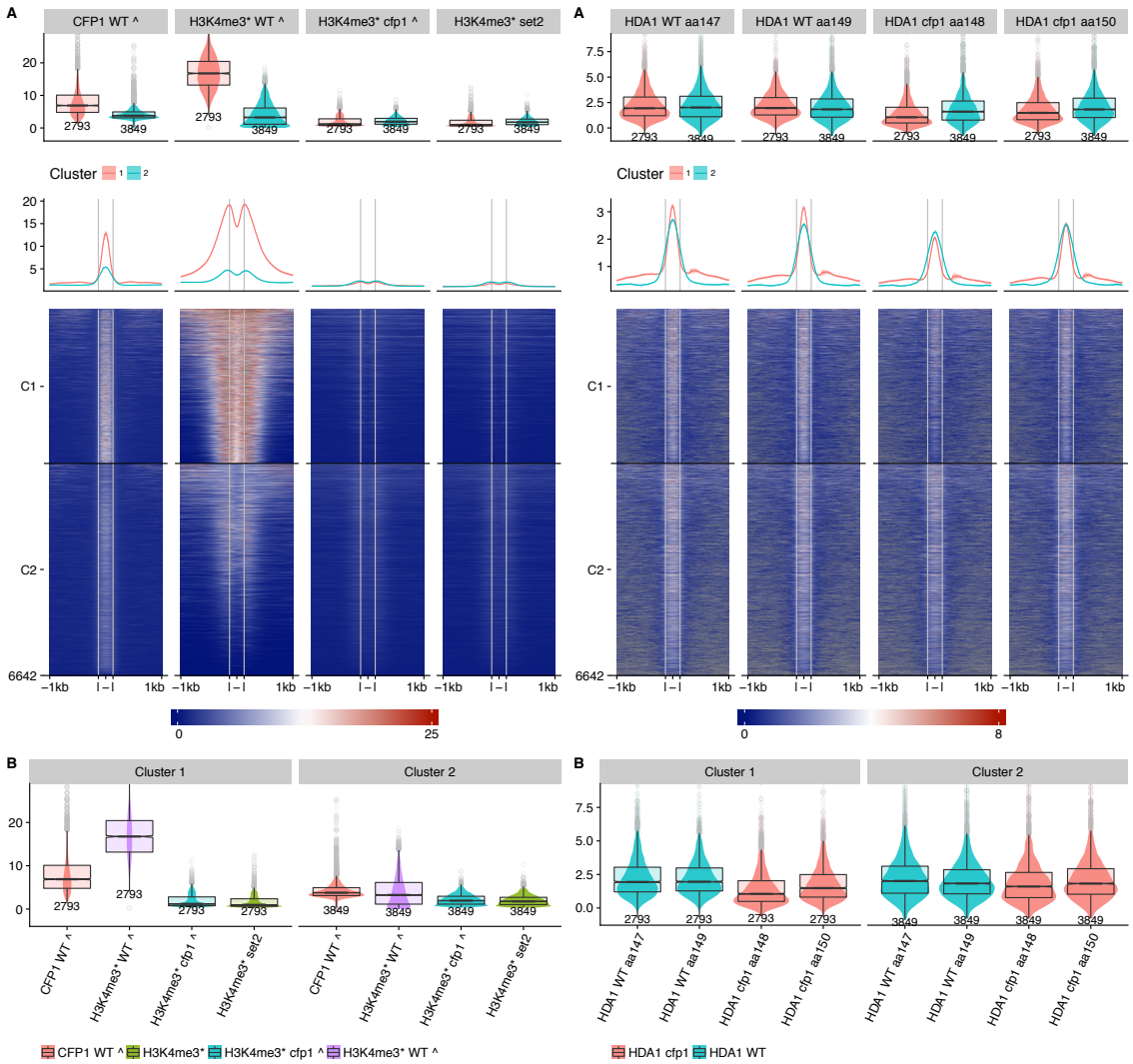


Figure 74 Quantification of HDA-1 in ChIP-seq in WT and *cfp-1* background shown in CFP-1 HiConf (C1) and LoConf (C2) peaks in context of CFP-1 signal and H3K3me3 in WT, *cfp-1* and *set-2* backgrounds. Panel A-left shows boxplot quantifications, profile plot quantifications and heatmaps for CFP-1 and H3K4me3 in WT, *cfp-1* and *set-2* mutant backgrounds. Panel A-right shows same plots for HDA-1 in WT and *cfp-1* strains. Panels B-left and B-right show same quantifications as boxplots on panel A, but are arranged in the way that makes it easy to compare replicates rather than clusters.

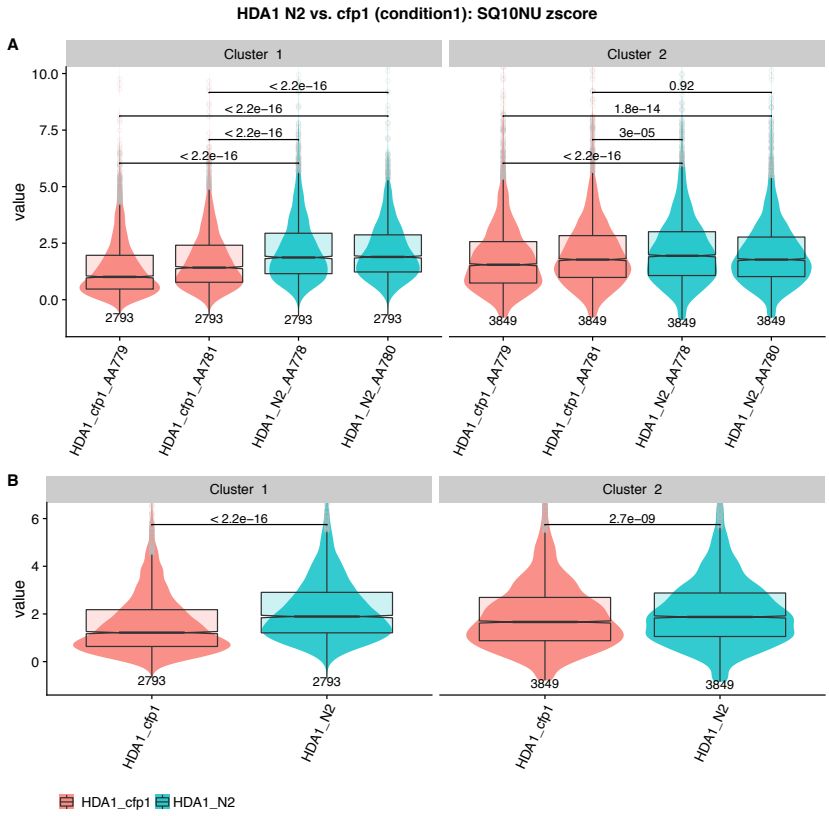


Figure 75 Significance testing for difference between HDA-1 in WT and *cfp-1* backgrounds. P-values are estimated using Mann-Whitney U test – a nonparametric method, which tests whether there is a location shift between two distributions (Bauer 1972).

3.21 Differential binding analyses in CFP-1 loci

Summarizing results from previous chapters I have found that:

- HDA-1 in *cfp-1* is significantly reduced in HiConf CFP-1 binding sites
- SIN-3 in *cfp-1* is significantly reduced in HiConf CFP-1 binding sites

Considering all uncertainty in above results, I conclude that a more stringent approach needs to be taken. The above analyses tested for the change in bulk of genomic loci and did not change if a change in given region is conserved between replicates. To achieve this goal, I analysed differential binding events in similar way I analysed differential expression in RNA-seq experiments.

I first tried to use DiffBind software (Ross-Innes *et al.* 2012), but I found this method has a design flaw that overestimated fold change in some regions, rendering whole

analyses not useful for more stringent analyses than using bulk data. DiffBind subtracts input signal from experiment, leaving the very small signal values in loci where experiment and input values are comparable (negative numbers are corrected to 1 read) – this leads to overestimation of fold change in such loci. This might have not been a substantial problem for *H. sapiens* analyses, where inputs are relatively shallow in comparison to ChIP-seq, but in *C. elegans* due to 30 times smaller genome size we have relatively deep input making this a substantial problem.

Another technical issue I have found during this analysis were towers (i.e. high signal regions produced by high numbers of copies of few unique reads) overlapping highly expressed RNAs. They were observed most likely due to "index hopping" phenomena (Kircher *et al.* 2012; Valk *et al.* 2018), where reads from multiplexed small RNA-seq library run in the same flow cell have contaminated ChIP-seq experiments. I removed this technical bias by filtering the assessed peaks to remove any overlaps with top 100 expressed RNAs, based on our RNA-seq data.

In brief, I developed a method based on HTSeq read counts in peak regions followed by DEseq2 differential analyses and fold change estimation. To asses if the differential binding is specific to assessed loci I inject randomised peak intervals into assessed loci. The number of randomised peaks is equal to the number of assessed peaks after filtering, and the chromosomal and width distribution is the same as original peaks. Start positions are drawn from uniform distribution ranging from 1 to chromosome length. If any randomised peak goes over allowed end of chromosome it's trimmed to this length. To make sure that randomised peaks are not assessing signal coming from actual peak regions or RNA towers, random peaks in +/- 1 kilobase around original peaks or top 100 RNAs are discarded. Combined filtered and random set constitute the set of loci used for tag counting, which is performed with the HTseq method. Than

counts are normalised and differential binding is called with DEseq2. Finally, stats from DEseq2 are retrieved and used to annotate assessed and random peak regions.

Group1	Group2	DB.DESeq2
HDA1_cfp1_m1	HDA1_N2_m1	764
HDA1_set2_m1	HDA1_N2_m1	16
HDA1_sin3_m2	HDA1_N2_m2	1
SIN3_cfp1_m1	SIN3_N2_m1	1091
SIN3_set2_m1	SIN3_N2_m1	13

Table 16 Number of differential binding events in all mutants.

First, I have assessed how many differential binding evens were found in non-random regions in each tested group. I have found a significant number of events in HDA-1 *cfp-1* and SIN-3 in *cfp1* vs matching WT differential binding calls. Other samples had very few events, which supports the observation that differential binding of HDA-1 in *set-2* and HDA-1 in *sin-3* replicates is disjoint. SIN-3 in *set-2* was an interesting case – here both replicates were significantly higher than WT ones in bulk, but hardly and differential binding events were detected. This might suggest that that there is an effect on SIN-3 binding in *set-2*, but it is not conserved between same loci in replicates. For further analyses I focused on HDA-1 and SIN-3 in *cfp-1*, as other mutants did not provide enough differential binding (DB) events to make any conclusive investigation.

Factor/strain	Class	# DB events	Mean FC	Median FC	Median FDR
All events					
HDA-1 in *cfp-1*	HiConf	2776	-0.27	-0.27	0.524
HDA-1 in *cfp-1*	LoConf	3799	0.01	0.03	0.876
HDA-1 in *cfp-1*	RND	5220	0.05	0.06	0.890
SIN-3 in *cfp-1*	HiConf	2776	-0.21	-0.22	0.296
SIN-3 in *cfp-1*	LoConf	3799	0.21	0.24	0.388
SIN-3 in *cfp-1*	RND	5220	-0.06	-0.05	0.563
Significant events					
HDA in *cfp1*	HiConf	234	-0.82	-0.76	0.011
HDA in *cfp1*	LoConf	9	-1.06	-0.95	0.001
HDA1 in *cfp1*	RND	0	-	-	-
SIN3 in *cfp1*	HiConf	549	-0.60	-0.69	0.014
SIN3 in *cfp1*	LoConf	370	0.62	0.65	0.020
SIN3 in *cfp1*	RND	24	0.80	0.85	0.009

Table 17 Summary of differential binding (DB) events in *cfp-1* background in CFP-1 bound and random (RND) loci. There are significant DBs in HiConf and LoConf for SIN-3 and mostly HiConf events for HDA-1.

I started with analysis of CFP-1 peak regions, divided into HiConf (2776), LoConf (3799) regions and injected randomised loci (RND, 5220). After performing the analyses, I found a number of significant events in both HiConf and LoConf CFP-1 regions for SIN-3 and mostly HiConf events for HDA-1 (**Table 17**). Further, I observed that HiConf events in both analysed factors represented almost exclusively a loss of factor binding, while in LoConf region SIN-3 DBs represented mostly gain of signal (**Figure 76**). I also observed that small number (9) of LoConf HDA-1 DB events represented loss of HDA-1 and small number of random SIN-3 events (24) represented almost only gain of SIN-3.

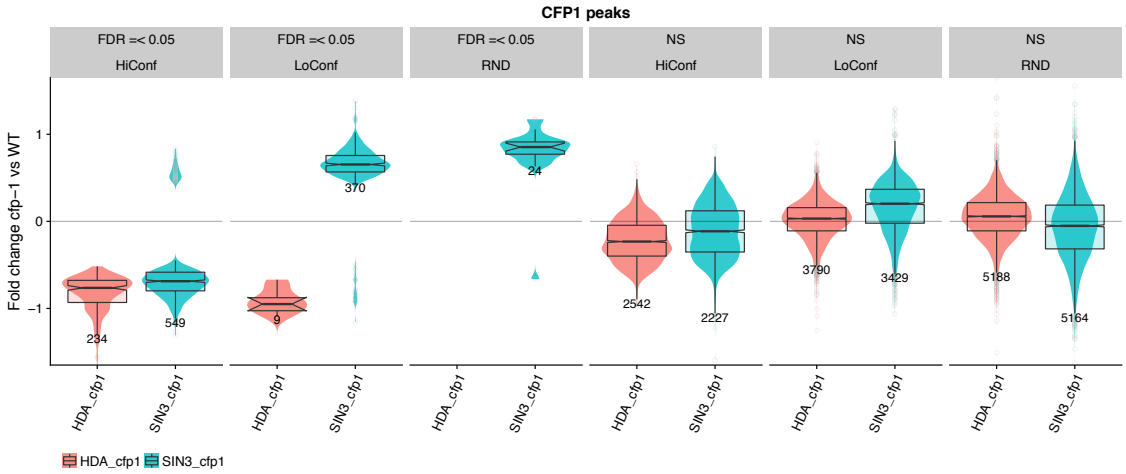


Figure 76 In *cfp-1* mutant HDA-1 and SIN-3 are reduced at HiConf CFP-1, but there is a gain of SIN-3 in low confidence CFP-1 loci. Boxplot shows the fold change of significant event (left) in context of non-significant events (right).

Taken together, this data allows me to speculate, that loss of *cfp-1* exclusively reduces HDA-1 occupancy. For SIN-3 the loss of *cfp-1* can both decrease or increase binding in context dependent manner – HiConf sites that loses H3K4me3 loses SIN-3 binding, while LoConf sites, where I detected no significant loss of H3K4me3 are gaining SIN-3 signal. The similar loss of H3K4me3 in *set-2* mutant allows further speculation, that CFP-1 has two distinctive modes of binding – it either binds with SET-2 (HiConf loci), and performs well characterised function of H3K4me3 deposition, or in other loci, where it can perform SET-2 independent function - mediating histone deacetylase complex activity. Hence, the cross talk with SET-2 and H3K4me3 might be crucial for CFP-1 role in SIN-3 recruitment – in H3K4me3/SET-2 occupied sites (HiConf) it can encourage SIN-3 deacetylation activity, while in other loci (LoConf) it can rather limit this activity. However, these observations can also support an alternative model. The assumption here is that SIN-3 is tethered to SET-2/COMPASS loci by CFP-1, and by different mechanism to other loci. When CFP-1 is not present, there is no competition for SIN-3 binding between SET-2/COMPASS (HiConf) and other (LoConf and RND) loci, so the excess of SIN-3 causes higher SIN-3 signal in other loci. This model is also supported by small number (24) of significantly increased binding events in randomly

sampled regions. This is in contrast to HDA-1, where no DB events in randomised regions were detected. This shows there is a potential of SIN-3 increase in loci not associated with CFP-1.

3.22 Differential binding analyses in SIN-3 and HDA-1 loci

Further I wanted to ask if loci associated with HDA-1 and SIN-3, rather than CFP-1, show similar differential binding events in CFP-1 mutant background. To answer this question, I have performed DB analyses with the same method as described in previous chapter but using HDA-1 and SIN-3 peak calls. To keep consistency with previous results I have divided HDA-1 and SIN-3 peaks into 3 classes – peaks overlapping with CFP-1 HiConf loci, peaks overlapping with CFP-1 LoConf loci and peaks not overlapping CFP-1. Also, randomly sampled peaks were injected.

Factor/strain	Class	# DB events	Mean FC	Median FC	Median FDR
All events					
HDA_cfp1	CFP1_cl1	2533	-0.32	-0.31	0.535
HDA_cfp1	CFP1_cl2	3060	-0.02	0.01	0.939
HDA_cfp1	non_cfp1	5327	0.03	0.04	0.909
HDA_cfp1	RND	8596	0.00	0.01	0.952
SIN3_cfp1	CFP1_cl1	2533	-0.22	-0.22	0.301
SIN3_cfp1	CFP1_cl2	3060	0.22	0.25	0.388
SIN3_cfp1	non_cfp1	5327	0.10	0.14	0.512
SIN3_cfp1	RND	8596	-0.10	-0.10	0.643
Significant events					
HDA_cfp1	CFP1_cl1	287	-0.82	-0.76	0.010
HDA_cfp1	CFP1_cl2	20	-0.82	-0.82	0.018
HDA_cfp1	non_cfp1	4	-0.47	-0.80	0.007
HDA_cfp1	RND	1	-1.27	-1.27	0.028
SIN3_cfp1	CFP1_cl1	552	-0.61	-0.68	0.013
SIN3_cfp1	CFP1_cl2	315	0.60	0.63	0.021
SIN3_cfp1	non_cfp1	148	0.52	0.70	0.025
SIN3_cfp1	RND	35	0.51	0.73	0.020

Table 18 Summary of differential binding (DB) events in *cfp-1* background in HDA-1 bound and random (RND) loci. There are significant DBs in HiConf, LoConf and not overlapping CFP-1 regions for SIN-3 and mostly HiConf events for HDA-1.

For HDA-1 I have annotated 2533 overlapping HiConf CFP-1, 3060 overlapping LoConf, 5327 not overlapping CFP-1 and 8596 random peaks. Similarly to CFP-1 peaks analyses, I observed high number of significant differential binding events on HDA-1 peaks overlapping HiConf regions for HDA-1 and SIN-3. HDA-1 peaks overlapping LoConf, non-overlapping CFP-1 and random regions had mostly SIN-3 DB events (**Table 18**). Also, in agreement with CFP-1 peak analyses, binding on HDA-1 peaks was decreased in HiConf regions for both HDA-1 and SIN-3, and increased for SIN-3 in other factors (**Figure 77**).

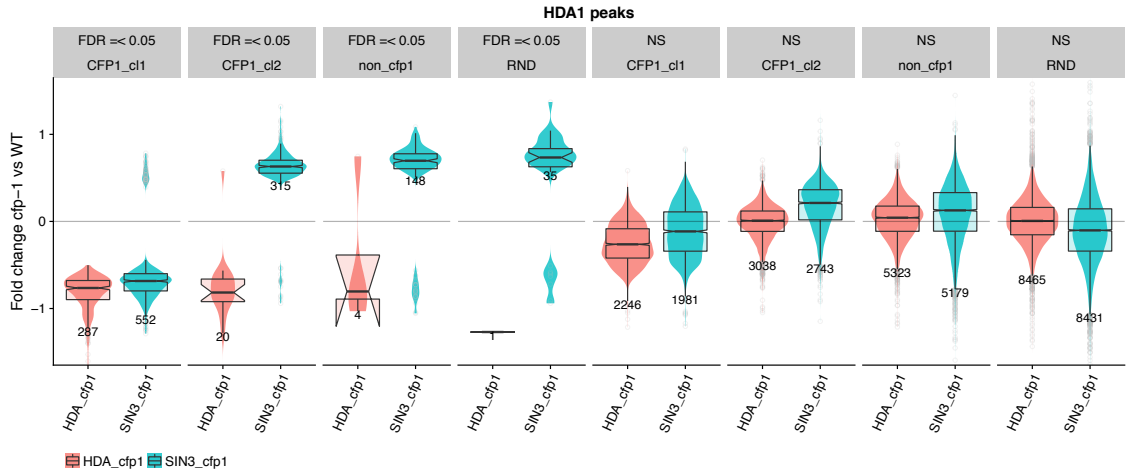


Figure 77 In *cfp-1* mutant HDA-1 and SIN-3 are reduced at HDA-1 peaks overlapping HiConf CFP-1, but there is a gain of SIN-3 in low confidence CFP-1 loci, HDA-1 peaks not overlapping CFP-1 peaks and random regions. Boxplot and violin plots show fold change of significant event (left) in context of non-significant events (right).

Factor/strain	Class	# DB events	Mean FC	Median FC	Median FDR
All events					
HDA_cfp1	CFP1_cl1	2451	-0.27	-0.26	0.454
HDA_cfp1	CFP1_cl2	2249	-0.01	0.02	0.851
HDA_cfp1	non_cfp1	1585	0.01	0.03	0.826
HDA_cfp1	RND	5363	0.06	0.06	0.855
SIN3_cfp1	CFP1_cl1	2451	-0.23	-0.24	0.210
SIN3_cfp1	CFP1_cl2	2249	0.23	0.27	0.320
SIN3_cfp1	non_cfp1	1585	0.18	0.20	0.396
SIN3_cfp1	RND	5363	-0.07	-0.07	0.653
Significant events					
HDA_cfp1	CFP1_cl1	302	-0.71	-0.67	0.014
HDA_cfp1	CFP1_cl2	20	-0.81	-0.70	0.018
HDA_cfp1	non_cfp1	2	-0.79	-0.79	0.018
HDA_cfp1	RND	3	-0.66	-1.17	0.001
SIN3_cfp1	CFP1_cl1	780	-0.55	-0.63	0.006
SIN3_cfp1	CFP1_cl2	413	0.55	0.59	0.014
SIN3_cfp1	non_cfp1	178	0.53	0.69	0.014
SIN3_cfp1	RND	50	0.51	0.72	0.020

Table 19 Summary of differential binding (DB) events in *cfp-1* background in SIN-3 bound and random (RND) loci. There are significant DBs in HiConf, LoConf and not overlapping CFP-1 regions for SIN-3 and mostly HiConf events for HDA-1.

I obtained very similar results when analysing SIN-3 peak calls. I divided SIN-3 peaks into 2451overlapping HiConf CFP-1, 2249 overlapping LoConf, 1585 not overlapping CFP-1, plus injected 5363 random peaks. I observed high number of significant differential binding events on SIN-3 peaks overlapping HiConf regions for HDA-1 and SIN-3, and additional significant DBs in SIN-3 peaks overlapping LoConf, SIN-3 peaks not overlapping CFP-1 binding sites, and random regions for SIN-3 only (**Table 19**). Also, SIN-3 and HDA-1 binding was decreased in HiConf overlapping regions, but SIN-3 was increased in all other DBs (**Figure 78**).

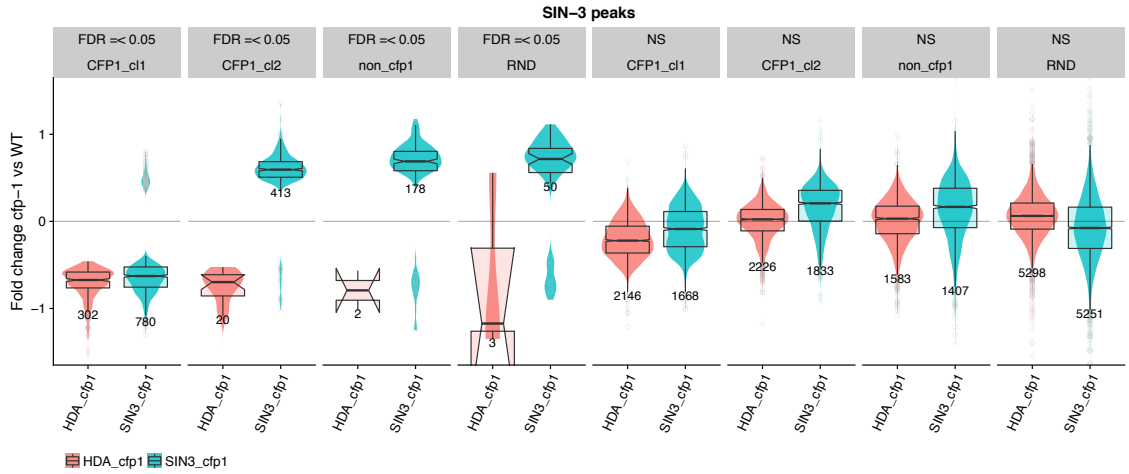


Figure 78 In *cfp-1* mutant HDA-1 and SIN-3 are reduced at SIN-3 peaks overlapping HiConf CFP-1, but there is a gain of SIN-3 in low confidence CFP-1 loci, SIN-3 peaks not overlapping CFP-1 peaks and random regions. Boxplot show fold change of significant event (left) in context of non-significant events (right).

Differential binding analyses for HDA-1 and SIN-3 strengthen previously proposed model, that loss of CFP-1 exclusively reduces HDA-1 occupancy, while for SIN-3 it reduces binding in loci where I also detected a reduction of H3K4me3 in *cfp-1* background and increases binding in all other loci. This effect is clear when analysing SIN-3 binding loci, where I detected almost 450 loci non-overlapping HiConf CFP-1 gaining CFP-1, in addition to few random loci loosing CFP-1 as well. This effect was even stronger when analysing DBs on SIN-3 loci – the number of non-overlapping HiConf CFP-1 DBs was close to 600. Interestingly, in addition to gains in SIN-3 loci, I also detected 50 SIN-3 differential binding events in random loci, with vast majority showing increased SIN-3. I observed no such effect for HDA-1, which suggests this is not due technical variability. This shows, that in *cfp-1* mutant background there is also significant gain of SIN-3 in loci not binding SIN-3 in wild type. In summary, though most DB events overlap with CFP-1 peaks, both in HDA-1 and SIN-3 peak analyses I have detected multiple CFP-1 gain of signal DBs in random regions not associated with CFP-1 peaks. Also, LoConf CFP-1 regions have significantly smaller CFP-1 occupancy, in comparison to HiConf peaks. This suggests a model, where SIN-3 gain of signal in *cfp-1* mutant is due to indirect, genetic interaction with CFP-1. Contrary, loss of both

SIN-3 and HDA-1 signal in *cfp-1* mutant happen exclusively in CFP-1 bound loci, and is most likely due to direct, physical interaction between CFP-1 and these proteins. To better understand connection between SIN-3, HDA-1 and CFP-1, and more generally crosstalk between H3K4me3 and histone acetylation further experiments are required. In particular, ChIP-seq for SIN-3, HDA-1 in *set-2* background, ChIP-seq for H3K27ac in *sin-3*, *hda-1* and *cfp-1* backgrounds and protein-protein interaction experiments between SIN-3, HDA-1 and CFP-1. However, in my opinion above analyses provide interesting hypothesis of how crosstalk between H3K27ac and H3K4me can be implemented in mechanistic way.

4 COMPUTATIONAL TOOLS, BIG DATA ANALYSES AND MACHINE LEARNING FOR GENOMICS

Genomics heavily relies on statistics, computational methods, and bioinformatics. In the previous two chapters I have shown how computational methods in biology facilitate hypothesis driven research on genomic scale. In this chapter I would like to focus on computational software I have developed to facilitate such research. I will first describe SeqPlots software for exploratory data analyses and visualization for genomics, and an rBEADS package for normalising ChIP-seq data. I will then introduce JADB – a user friendly database and collection of computational pipelines and auxiliary tools for annotating, cataloguing, fetching and analysing ChIP-seq and RNA-seq data. Finally, I would like to go beyond hypothesis driven research and show how advanced computational methods can facilitate data driven research, where without any prior assumption one can extract useful knowledge from big biological datasets.

4.1 Computational tools overview

To support genome wide investigation of expression regulation in the context of chromatin factor binding and histone modifications I developed a number of software

tools and methods. These tools can be classified into 3 categories: (1) databases and tools for automatic data recovery, (2) automated quality control and data processing, (3) tools for statistical analyses and visualization of genomic data. The most important tools I created are:

- **JADB** is in-house database and collection of pipelines, which allows efficient processing, storage and retrieval of our vast data collection: 1300 ChIP-seq experiments and 311 RNA-seq experiments, all processed using automated pipelines.
- **JADBtools** is an R package that provides JADB with an efficient data retrieving mechanism and a bridge to interact with downstream software.
- **rBEADS** is an R package that enables ChIP-seq data normalization
- **DEview** is a interactive tool that fetches the RNA-seq data from JADB and performs differential expression analyses and gene ontology enrichment. It is especially useful to analyse and visualise time course transcriptomics.
- **CorView** is a tool for plotting interactive heatmaps based on track correlation, plus performing Gaussian graphical models and standard factor analysis on tracks in JADB using JADBtools API.
- **SeqPlots** is exploratory data analysis (EDA) and visualization package, available on Bioconductor and as stand-alone desktop application.
- **ARD** – automated replicate detection – a tool that can automatically select the best replicates based on a number of metrics - correlation based quality metrics, peak call statistics and various quality checks. It also combines peak calls using irreproducible discovery rate (IDR) and intersection methods, combining ChIP replicates using optimised IDR method

I also worked on the non-parametric sparse factor analysis (NSFA) algorithm. I implemented NSFA method in R, which gives three advantages: (1) more

computationally efficient; (2) compatible with normalization (rBEADS) data acquisition (JADBtools), and visualization (SeqPlots) packages I have authored; (3) R is open and free environment, easy to deploy on our computing cluster, more accessible to scientific community. Also, I assured numerical consistency between Matlab and R implementations – this was not a trivial problem as it required to generate same random numbers in both environments. Additional work on NSFA included:

- R package for simulating ChIP-seq tracks *in silico* based on Gaussian random process
- Calculation of stats on BigWig files, including coefficient of variance, particularly useful, to evaluate if NSFA will perform well on a given dataset.

Research to assess the impact of de-duplication of ChIP-seq experiments: this led to a conclusion, that 1 read/loci cutoff is too stringent for narrowly distributed, high coverage sequencing data, such as RNA polymerase ChIP-seq experiments.

4.2 SeqPlots - Interactive software for exploratory data analyses, pattern discovery and visualization in genomics

As shown in Chapters 2 and 3, practical, genome-wide analyses of genomics datasets requires the analysis of a large number of genomic loci across numerous experiments. To enable performing such analyses easily and quickly, both by me and those who lack bioinformatics training I have created SeqPlots software. This section describing SeqPlots contains text published in (Stempor & Ahringer 2016). SeqPlots is available at http://bioconductor.org/packages/seqplots and https://github.com/Przemol/seqplots.

4.2.1 Introduction to SeqPlots software

In brief, sequencing based techniques, such as ChIP-seq and RNA-seq are widespread experimental tools that generate vast amounts of data for downstream analyses such as uncovering global patterns of genomic activity. After aligning sequence reads to the

reference genome, read coverage is calculated. Visualizing coverage tracks using genome browsers is the simplest way to inspect the results. Nevertheless, calculating and plotting signals across groups of selected genomic locations is essential for genome-wide hypothesis testing and quantitative comparisons.

Typically, users plot the abundance of signal (e.g., read coverage) across a set of genomic regions (e.g., transcription start sites) either as a profile plot of average signal or as stacked rows of individual signals visualized as a heatmap. Such plots are usually generated using online or command line tools such as Galaxy/Cistrome, ngs.plot, and deeptools, or using custom scripts combined with plotting software such as Gnuplot (Huber *et al.* 2015; Liu *et al.* 2011a; Ramírez *et al.* 2014; Shen *et al.* 2014; Williams *et al.* 2013). I found these methods were either laborious, as each plot needed to be set up individually, or were difficult to use by those with little computational training. These can discourage users from generating a large number of plots for data exploration.
To address this, I developed SeqPlots, a highly configurable, graphical user interface (GUI) operated application that rapidly generates publication quality average profile plots or heatmaps that can be clustered using different algorithms to uncover patterns within the data. A key feature of SeqPlots is the ability to select a set of features and signals, then rapidly plot them in any combination, facilitating wide data exploration.

4.2.2 SeqPlots capabilities

SeqPlots can plot signals from any experimental or *in silico* data (e.g. ChIP-seq or RNA-seq read coverage, density of sequence motifs, mappability, nucleosome occupancy) over one or multiple sets of genomic features, (e.g. TSSs, gene bodies, peak calls). Users first add signal tracks and genomic feature files to an integrated SeqPlots database (see **Table 20** for accepted file formats). Then any combination of signal and feature files in the database, together with any user entered sequence motifs, can be

analyzed. Plots can be anchored at either end of a feature, at both ends, or at centers, and users can define which lengths of upstream and downstream sequence to plot. Additionally, three different methods can be used to cluster heatmaps: k-means, hierarchical clustering, and self-organizing maps (unsupervised neural networks); heatmap rows can also be sorted by signal strength.

File formats	Recognized extensions	Additional notes
Genomic feature formats		
General Feature Format	gff or gff.gz	Can be compressed with gzip
Browser Extensible Data	bed or bed.gz	Can be compressed with gzip
General Transfer Format	gtf or gtf.gz	Can be compressed with gzip
Signal track formats		
bigWig Track Format	bw	Preferred track format
Wiggle Track Format	wig or wig.gz	Converted to bigWig upon upload, can be compressed with gzip
BedGraph Track Format	bdg, bdg.gz, bedGraph or bedGraph.gz	Converted to bigWig upon upload can be compressed with gzip
Binary Sequence Alignment/Map	bam	Coverage is calculated using all aligned reads

Table 20 File formats supported by SeqPlots

4.2.3 SeqPlots implementation

SeqPlots utilizes indexing and the multi-layer summarization properties of bigWig files for rapid data acquisition (Kent *et al.* 2010), and precalculates and stores profiles for all combinations of selected signals and features. Users are presented with a clickable array of signal/feature pairs that can be plotted individually or in any combination in a matter of seconds. Average profile plots or heatmaps are immediately displayed as previews and can be downloaded as PDF files. Profile plots can display standard error and 95% confidence intervals. Spreadsheets with annotated heatmap clusters can be downloaded for downstream analyses such as additional clustering or gene enrichment analyses. Scaling, colors, axes, and titles are also easily configurable. Signal and feature files uploaded to the integrated SeqPlots database are available for use in later plot setups.

Users can search and sort uploaded files, and annotate them with comments, user names and reference genome versions.

4.2.4 Example of SeqPlots analyses

Figure 79 illustrates a typical use of SeqPlots. Five feature files in bed format containing genomic coordinates of protein coding genes in different expression bins were selected together with three bigWig signal files (normalized read coverage of H3K4me3, H2A.Z, and H3K36me3). In addition, the dinucleotide motif CG was inputted and SeqPlots generated a CG density track for use in the analyses. A plot type anchored at the start position (the TSS) was then selected, and 1 kb upstream and 1.5 kb downstream of the TSS was specified. Following the setup and calculation, SeqPlots presented a clickable grid (**Figure 79** a and b, top). Selecting the desired combinations and plot type (average profile plot or heatmap) generates a plot. In **Figure 79**a, three signals (H3K36me3, H3K4me3, and H2A.Z) and one feature (top 20% TSSs) were selected for an average profile plot. For **Figure 79**b, these were deselected and a new combination was selected (H3K4me3 and all five TSS expression classes). For **Figure 79** c and d, single combinations of feature and signal were selected.

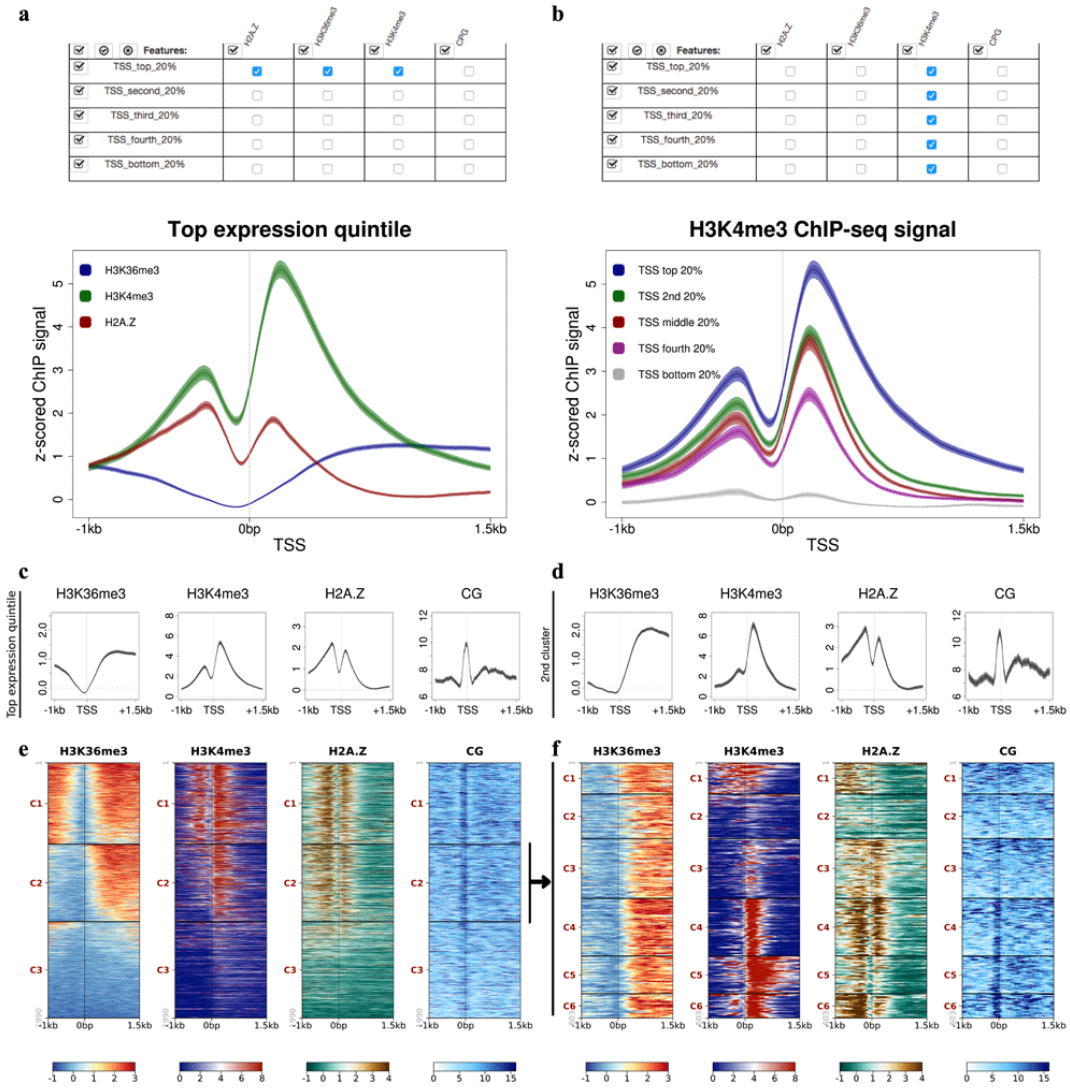


Figure 79 An example of SeqPlots workflow to analyse H2A.Z, H3K36me3, H3K4me3 and CpG density across C. elegans protein coding TSSs separated by expression quintiles (a, b). Top, GUI interface showing clickable grid of signal/feature combinations. Bottom, plots resulting from the clicked selections. (c) Plots of individual signals across genes in top expression quintile anchored at TSSs, plotting 1 kb upstream and 1.5 kb downstream of TSSs, and (e) heatmaps generated using k-means clustering (3 clusters) of TSSs in top expression quintile, using H3K36me3 signal for clustering. (d) Average signal profiles and (f) heatmaps generated from cluster 2 (C2) in (e) made by downloading full cluster data and uploading file with cluster 2 regions. Heatmaps were clustered using H3K4me3, H2A.Z and CpG signals. Data used to generate this figure are available from GEO (H3K4me3: GSE28770 - https://www.be-md.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28770; H3K36me3: GSE62833 - https://www.be-md.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62833; H2A.Z/HTZ-1: GSE49717 - https://www.be-md.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49717). TSS annotations (Chen *et al.* 2013; Kruesi *et al.* 2013), or Wormbase/Ensembl 81 if a gene had no TSS annotation in either dataset (available from https://gist.github.com/Przemol/c5114067cc2dd236ed1dbcaf41003472). Genes were divided into expression bins using DCPM values (Gerstein *et al.* 2014)

A three-cluster heatmap was then generated using all four signal tracks, clustered using just the H3K36me3 signal. This simple clustering identified regions with bidirectional (C1), unidirectional (C2) or little (C3) H3K36me3. The unidirectional cluster (C2) was extracted from the cluster annotation spreadsheet and uploaded for further re-clustering. A self organizing map with 6 neurons was applied to the three other features – H3K4me3, H2A.Z and CpG, revealing clusters with different patterns of H3K4me3 and H2A.Z marking. For example, cluster C4 shows strong H3K4me3 downstream of the TSS and H2A.Z enrichment both upstream and downstream of the TSS whereas cluster C6 has a similar H3K4me3 pattern, but H2A.Z shows higher enrichment upstream of the TSS. Additionally, clusters C4–C6 have a stronger CpG signal at the TSS than clusters C1–C3. This simple example shows how SeqPlots can be used to find relationships between genomic features and signals. The rapid plotting capability and ease of use of SeqPlots should facilitate wide exploration of high-throughput sequencing data, leading to the discovery of novel biological associations.

4.2.5 Software availability

SeqPlots is distributed as user-friendly stand-alone applications for Mac, Windows and Linux, and is available as an R programming language package from the Bioconductor repository. SeqPlots can be also deployed as a server application, which is useful for data sharing within laboratories, collaborative usage and remote work. SeqPlots is an open source and open development project: source code wiki, bug tracker and pull requests are available via GitHub. SeqPlots is licenced as LGPL 2.1 (*https://www.gnu.org/licenses/old-licenses/lgpl-2.1.html*). Software is available from:

- *http://przemol.github.io/seqplots* (Mac, Windows, Linux, full documentation)
- http://bioconductor.org/packages/seqplots (R/Bioconductor)
- http://przemol.github.io/seqplots/#installation---server-deployment (server deployment)

- *https://github.com/Przemol/seqplots* (latest source code, open development tools, including wiki, bug tracker, and pull requests)

4.3 rBEADS – R implementation of Bias Elimination Algorithm for Deep Sequencing

ChIP-seq experiments can be influenced by multiple unwanted factors including nucleotide composition bias, mappability variations and differential local DNA structural effects. For example, GC-rich sequences are often over-represented, while AT-rich sequences are depleted. Also, the ability to confidently map reads to the reference genome (termed mappability) varies due to duplicated and low complexity regions. Finally, local DNA or chromatin structural effects can lead to inhomogeneity of sequencing data coverage. To address this problem the BEADS algorithm was developed in our laboratory (Cheung et al, 2011). The algorithm involves 3 steps – GC correction, masking of low mappability regions and normalization to input. The original BEADS algorithm in Python/Java implementation (http://beads.sourceforge.net/) was not suitable for long term development or ease of usage by external users. The code was also not documented, and the pipeline relied on many external tools. For these reasons I decided to re-implement the algorithm. R version simplifies usage of the software, as well as introduces improvements and fixes issues identified since original release. Key differences from original implementation are:

- Introduction of BEADS score – a final metric of the signal, that is normalised for depth of coverage of both input and ChIP experiment and have intuitive value of 1 being noise level in ChIP and higher values being enrichment over this value, while lower being depletion. It scales well in both log2 and z-score.
- Enriched region finder algorithm, which removes a need to use external peak caller in GC normalisation step.

- Chromosome by chromosome GC and input analyses, that can easily be parallelized for better performance and allows to run rBEADS on laptop class computers for big mammalian genomes.
- It works with GEM mappability tracks, which removes long and error prone process of simulating genomic reads and aligning them with aligner software. It also removes a need for an external aligner, as long as source data are provided in aligned BAM format.
- It works directly with binary files – takes BAM files as input and produces BigWig files as output – this saves computing time, memory requirements and disk storage.
- It provides methods to use summed input – a meta-input track created from multiple input controls for ChIP experiments – this helps with the issue of artefacts with shallow inputs and improve consistency of normalisation procedure.
- The GC correction step was revised – I found very high or low GC content regions to introduce artefacts in normalisation. I implemented an automated filter, that identifies regions with extremely high or low GC and marks them as unadaptable.

The rBEADS package is publicly available (Stempor 2014). For implementation details please see https://github.com/przemol/rbeads/wiki/Implementation.

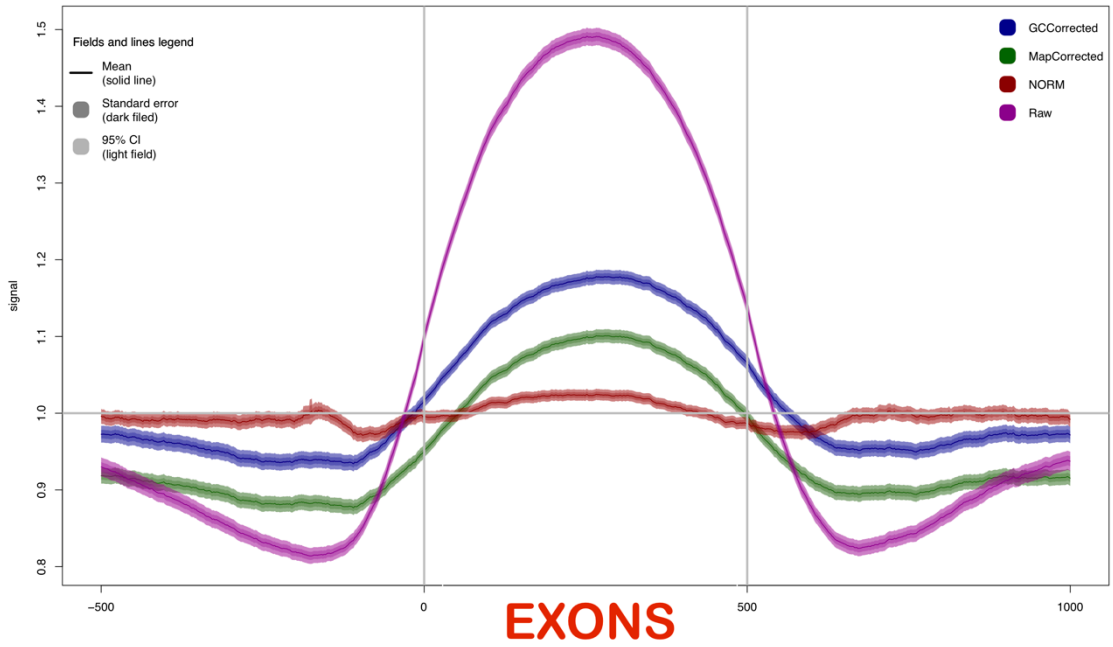


Figure 80 The three steps of rBEADS normalisation on validation plot. This plot shows normalising input track over exonic regions. Since in input experiment no immunoprecipitation has been performed, we expect input to be generally flat. However, *C. elegans* exons have high GC content relative to the genome. Raw sequencing data has a GC bias which shows input enrichment at exons. The plot show 3 steps to correct it - in violet is the original signal, in blue the GC normalised signal, in green the signal with additional non-mappable regions filtering, and in red after final steps of beads – division. The signal is z-scored to compare enrichments that are in different scales, e. g. raw reads vs. BEADS score. In order to compare signal coming from different experiments y-axis show z-scored value of read count, intermediate rBEADS step scores and final rBEADS score.

4.3.1 General rBEADS pipeline structure

In this section I briefly describe the steps of rBEADS, then in the next section I explain them in more detail. The pipeline starts with importing BAM or Eland formatted alignment files into R environment (**Figure 81**). The reads are filtered to permit only the ones with sufficient mapping quality. By default, only the reads above an alignment quality threshold of 10 are permitted both for BWA and Solexa pipeline alignments. Than the reads are extended to the expected fragment length (200bp by default).

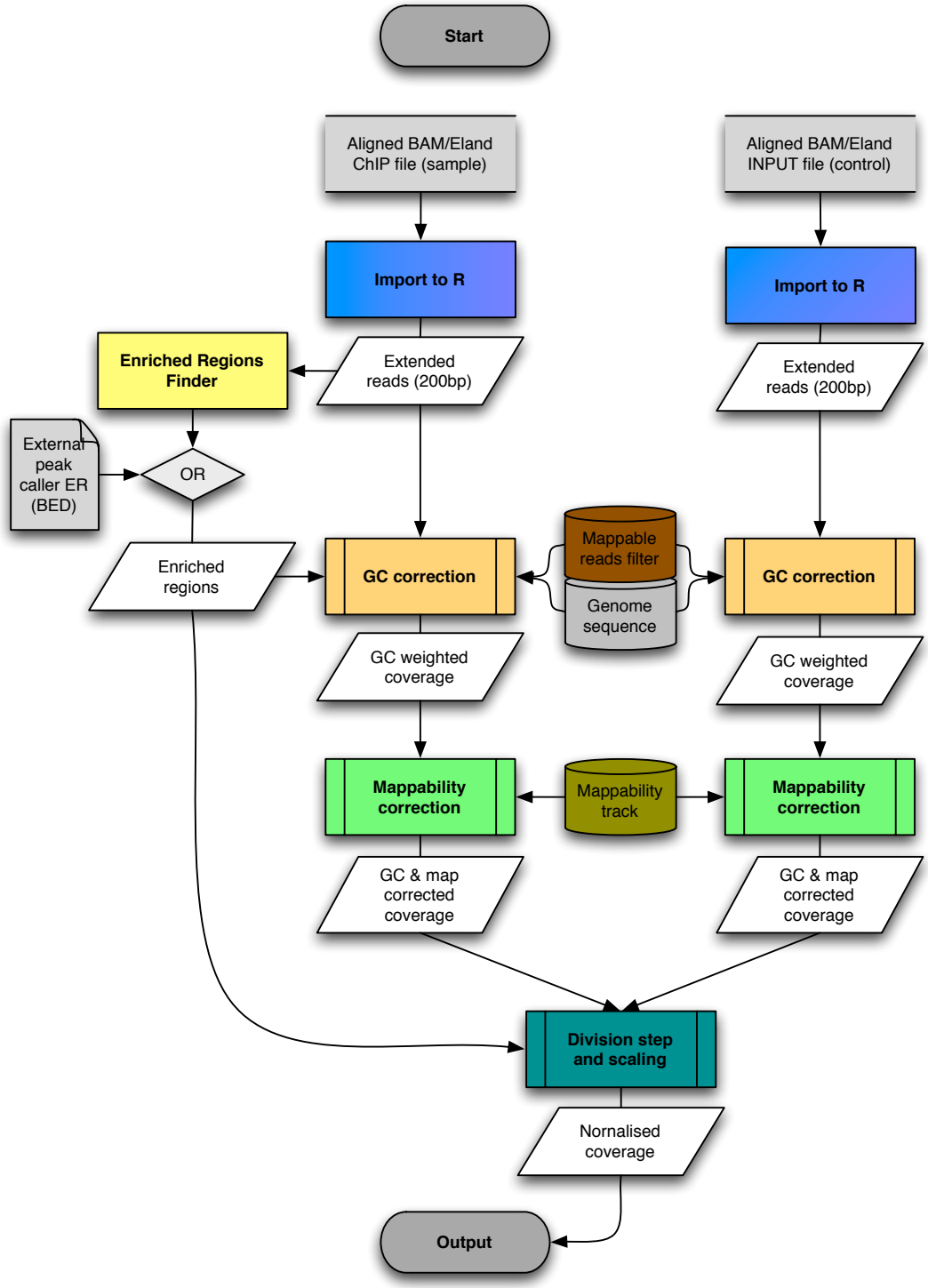


Figure 81 Summary of rBEADS pipeline, taking aligned files as input and outputting normalised coverage files.

The further optional step involves finding enriched regions using reads imported in the first step. Alternatively, any peak calling software (e.g. MACS: http://liulab.dfci.harvard.edu/MACS/) can be run on the input alignment file and resultant BED file might be utilized to identify enriched regions.

Extended read and enriched region information are used to correct GC bias. This step also utilizes mappable reads filter and genomic sequence assembly - both provided with a package. The output of this step is GC corrected read coverage track.

The next step masks non-mappable regions. The mappability tracks for major model organisms are provided. They can be also calculated using GEM-mappabaility software from a FASTA file or reference genome package in R. This step outputs GC bias and mappability corrected reads coverage.

The above procedure is also applied to the "*Input*" sequence obtained from sequencing DNA in the starting extract (no immunoprecipitation). The only difference is that enriched region masking in the GC bias correction step is not done - the *Input* track is expected not to have any enriched regions. If no *Input* specific for the run is available, summed inputs from other experiments (EGS or formaldehyde) provided with the R package can be used.

The final step of rBEADS pipeline includes division of ChIP coverage by run specific or summed Input coverage and scaling the output to achieve fully normalised BEADS scores. After this step the output may be further z-scored or log2 scaled and saved as BigWiggle track.

4.3.2 Import BAM formatted alignment files

In this step BAM file is scanned using binary flag, which allows only correctly mapped reads to be imported (**Figure 82**).

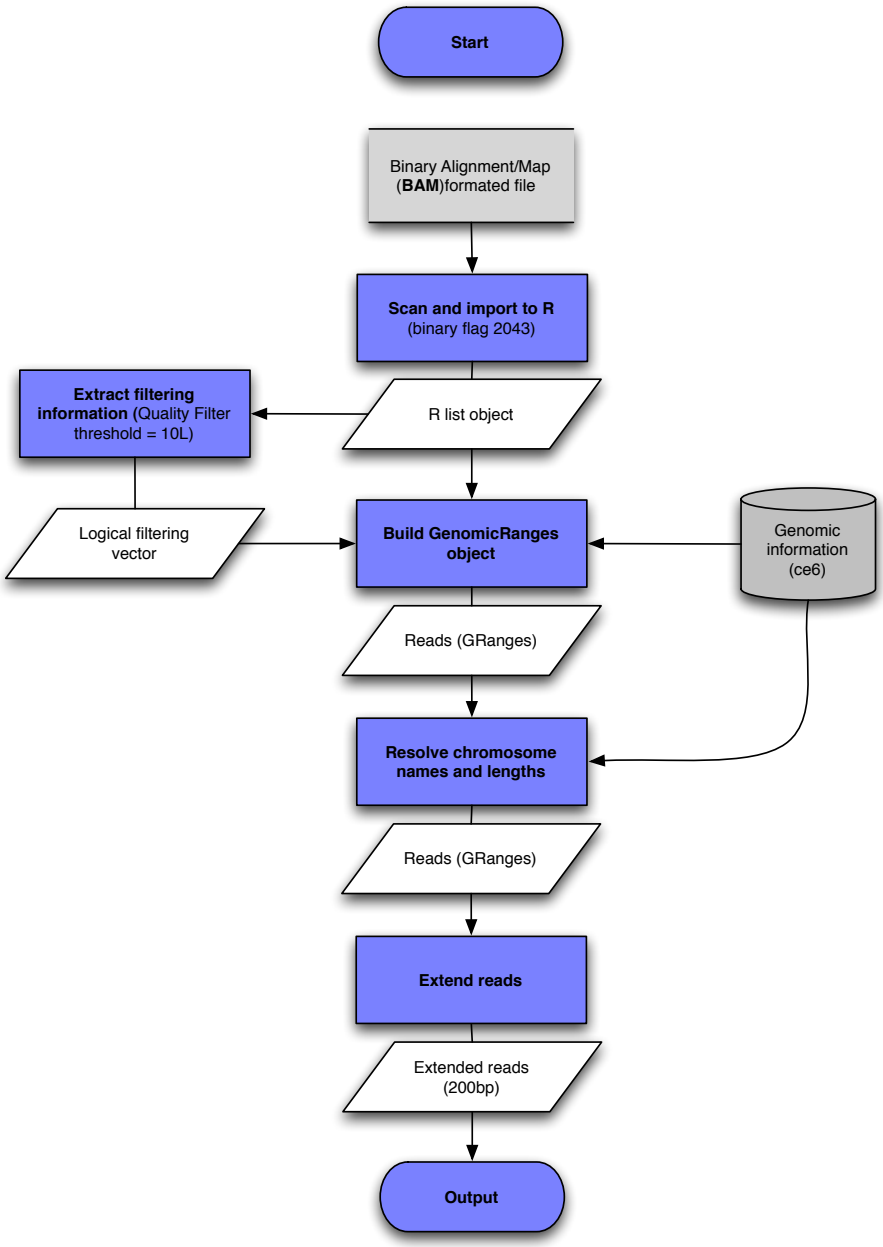


Figure 82 Flowchart presenting the procedure for importing aligned file (BAM), filtering, extending and creating coverage track.

The R list object is created from imported reads. Subsequently, this object is filtered to contain the reads with BWA quality score higher than 10 and "GRanges" object (defined in "GenomicRanges" package) is constructed. At the end of this step the chromosomes' names and length information are attached to GRanges object. Reads are extended to expected length (200bp by default). Optionally the GRanges object can be uniqued (filtered in a way that only unique reads are retained) in order to discard reads multiplied by PCR amplification and sequencing scanner errors.

4.3.3 The preparation of summed Input from multiple Input BAM files - sequencing depth adjustment and summarization

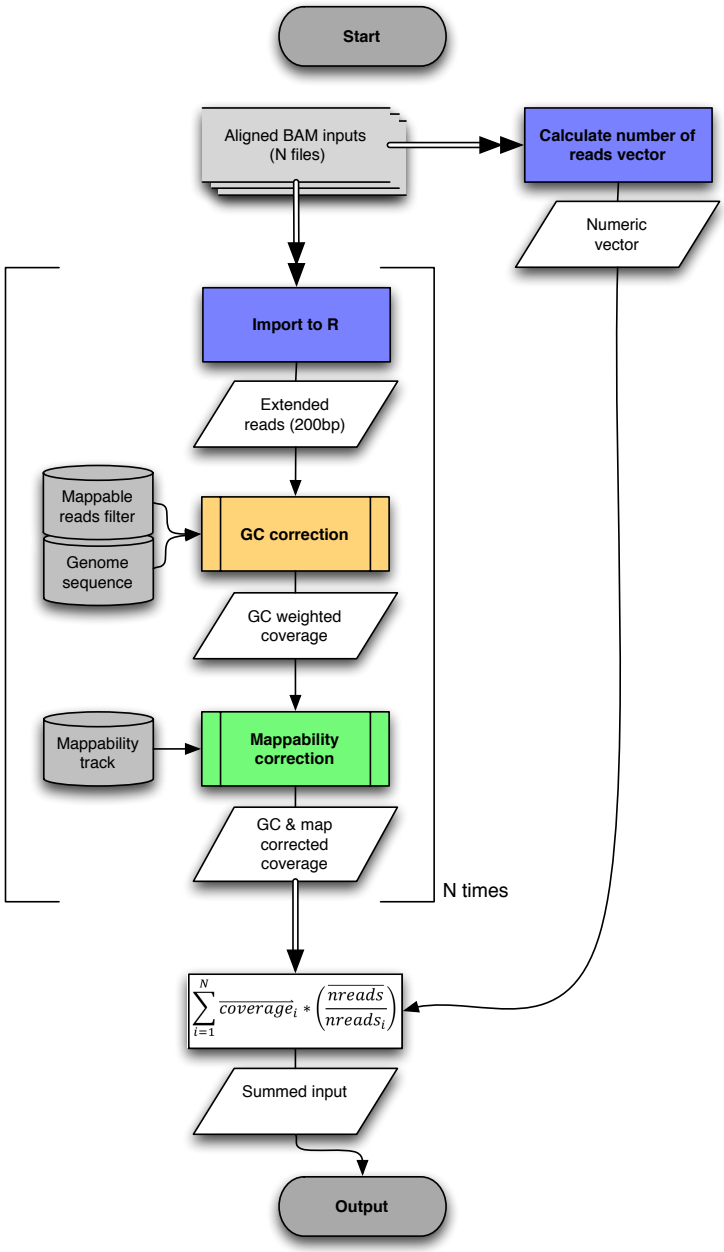


Figure 83 Meta-inputs – flowchart showing preparation of summed input from multiple input BAM files, sequencing depth adjustment and summarization.

The presented procedure is implemented in "SumBAMinputs" function included in rBEADS package and should be used to obtain summed input in R binary format. The algorithm takes as input multiple BAM files (**Figure 83**). The first two steps of BEADS normalization (GC correction and non-mappable regions masking) are separately performed on every Input file. Since the files are ChIP *Inputs* (sequenced library

without performing immunoprecipitation) the enriched region step is omitted. Subsequently each coverage track is scaled by the ratio of mean reads' number in all Input BAM files to file specific number of reads in order to equalize tracks with different sequencing depths. Finally, all tracks are summed, and range encoded list structure is saved to R binary file. This file can be identified as summed input in main function in rBEADS package.

4.3.4 Automatic enriched regions (ER) detection

This simple algorithm has been designed to eliminate the requirement of using external peak calling software proposed in original BEADS implementation and therefore simplifies and automates rBEADS pipeline (**Figure 84**). However, external peak callers might still be used for more precise ER detection. In case of providing the rBEADS with external BED file this step will be omitted. The algorithm starts with calculating the coverage from extended reads. Subsequently this coverage is sliced at 75-th percentile of coverage values creating potentially enriched islands. The area under the curve for each island is calculated (integer sum of the coverage in potentially enriched islands). The top 10% islands (those with the highest field values) are considered as enriched regions. In the final step the "GRanges" object containing enriched regions is created and utilised in further steps. Additionally, the BED file of enriched regions is exported.

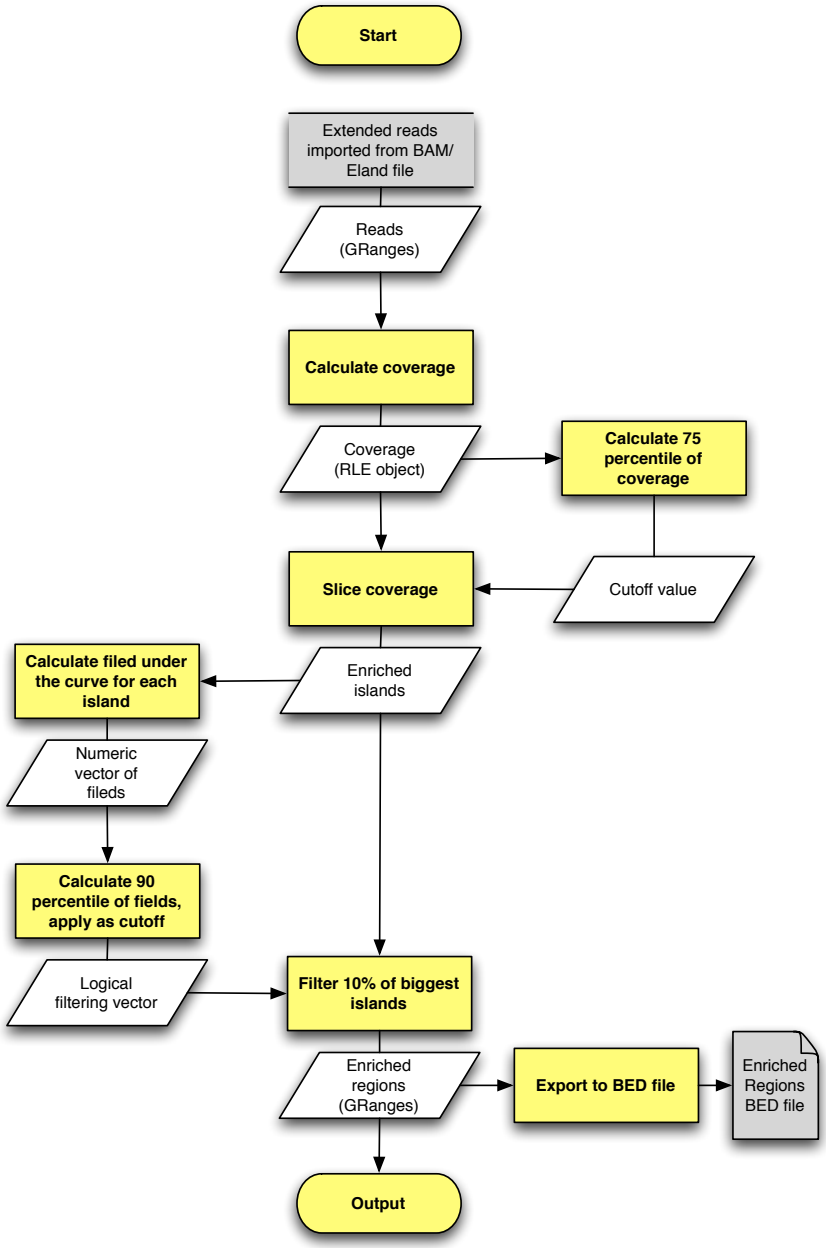


Figure 84 Enriched region detection algorithm used for rBEADS GC content normalisation step.

4.3.5 The GC bias correction

The GC bias correction step (**Figure 85** and **Figure 86**) takes as input following structures: (1) extended reads stored in "GRanges" object, (2) enriched regions stored in BED file or "GRanges" object, (3) mappable reads filter and (4) genome assembly.

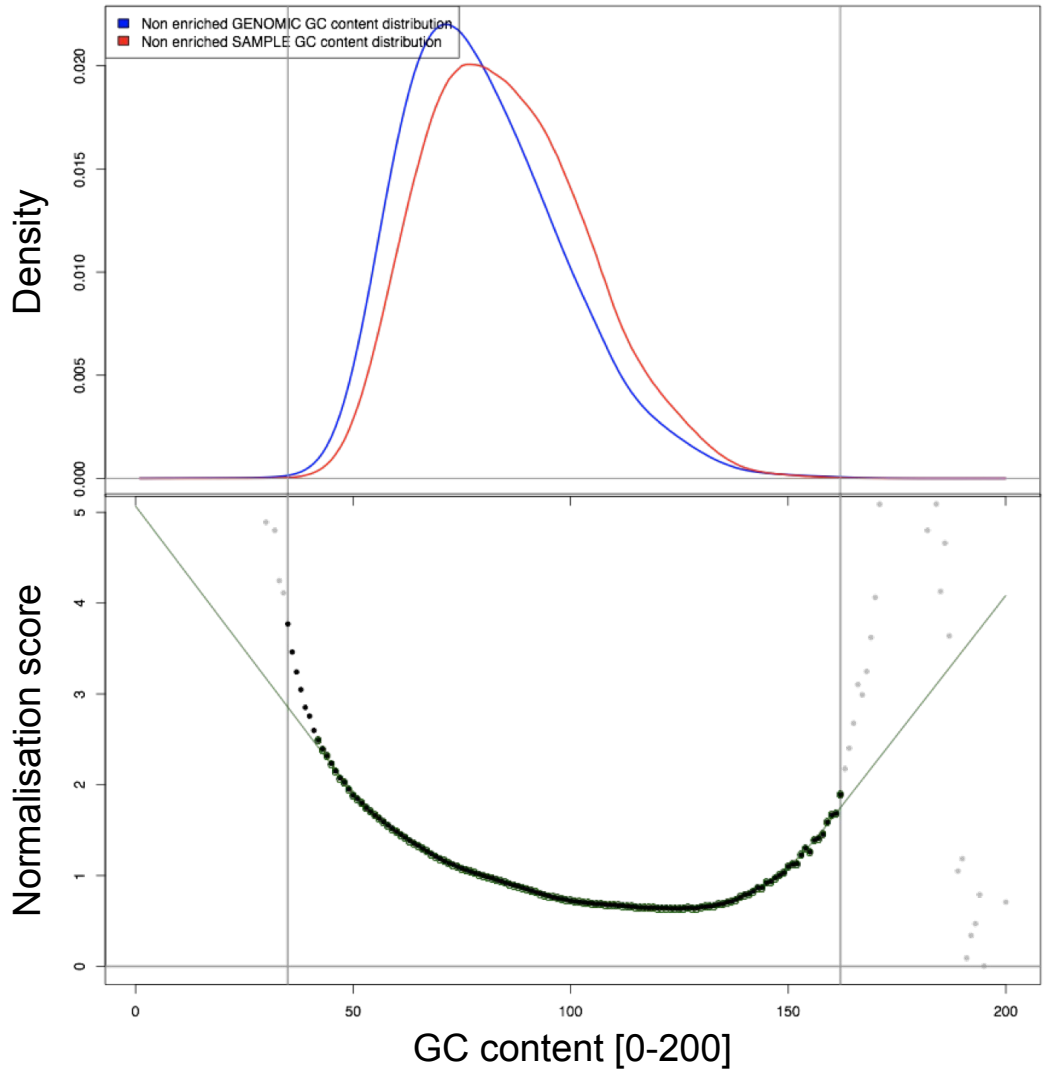


Figure 85 Principle of GC normalisation with rBEADS. Upper plot show GC content density of probability for *C. elegans* genome (blue) and for uniform, non-enriched reads from Illumnain sequencing platform. After dividing genomic distribution by reads distribution I obtain normalisation scores. Each read is assigned proper normalisation score based on individual GC content. This function is visualised on lower panel. Finally, a weighted coverage is calculated based on normalisation scores.

For each read GC content is calculated (for default read length of 200bp the GC-content is integer value between 0 and 200). Subsequently two histograms are calculated – first represents GC-content distribution for reads in non-enriched regions while the second represents genomic GC distribution in non-enriched and mappable regions of the genome (**Figure 85**). These histograms are divided - counts in reads distribution are divided by counts in genomic distribution - creating the scores for each GC bin. For expected read length of 200 the vector of 201 scores is created representing all possible

GC-contents in reads. Then each read is assigned a score depending on its GC-content. This creates the weighted vector for the coverage calculation. In the final step non-GC-correctable regions are masked out from calculated coverage. The range length encoded list (the "SimpleRleList" object containing coverage for every chromosome) is returned and used in the following step.

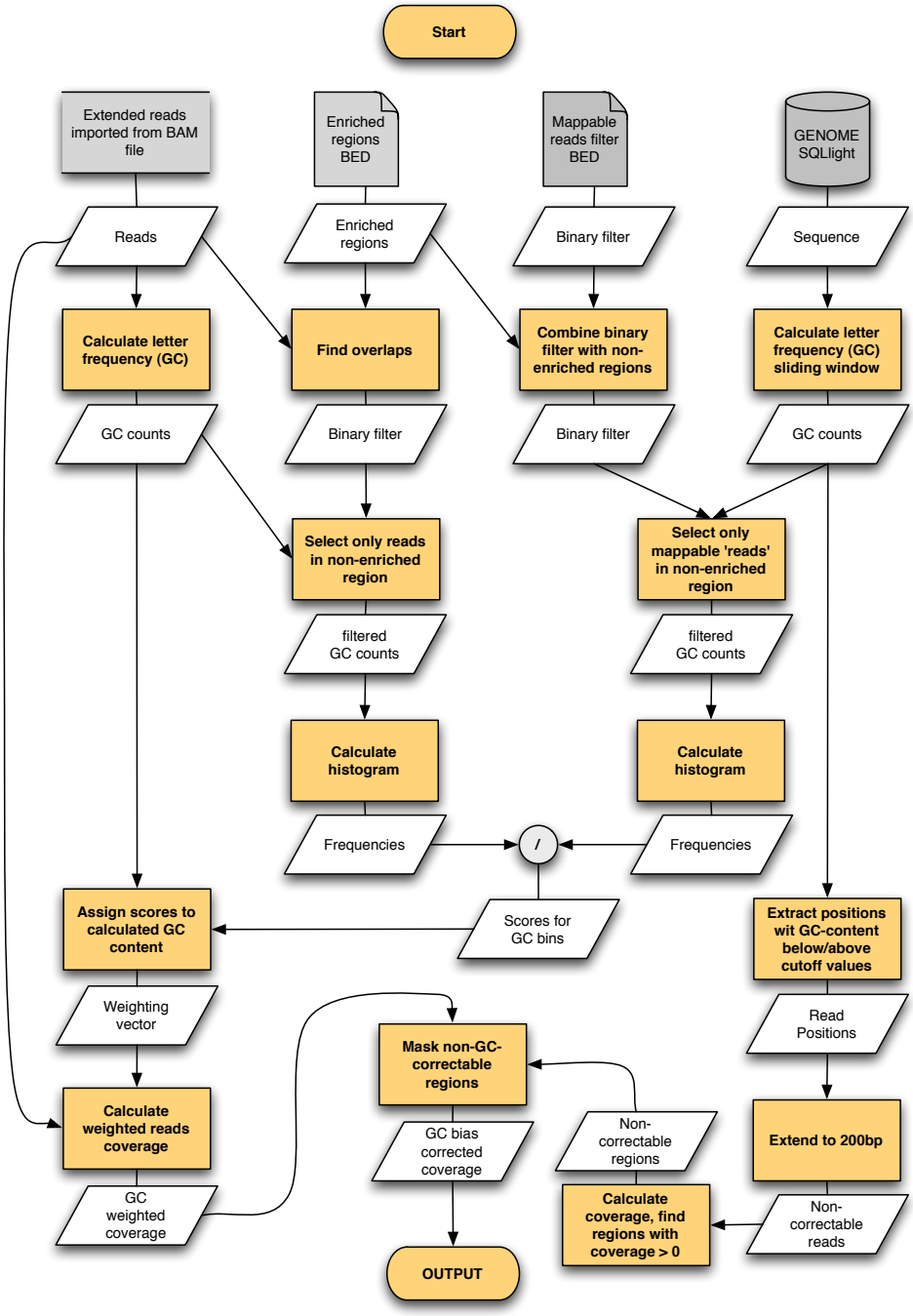


Figure 86 GC bias correction flowchart

4.3.6 Non-mappable regions masking

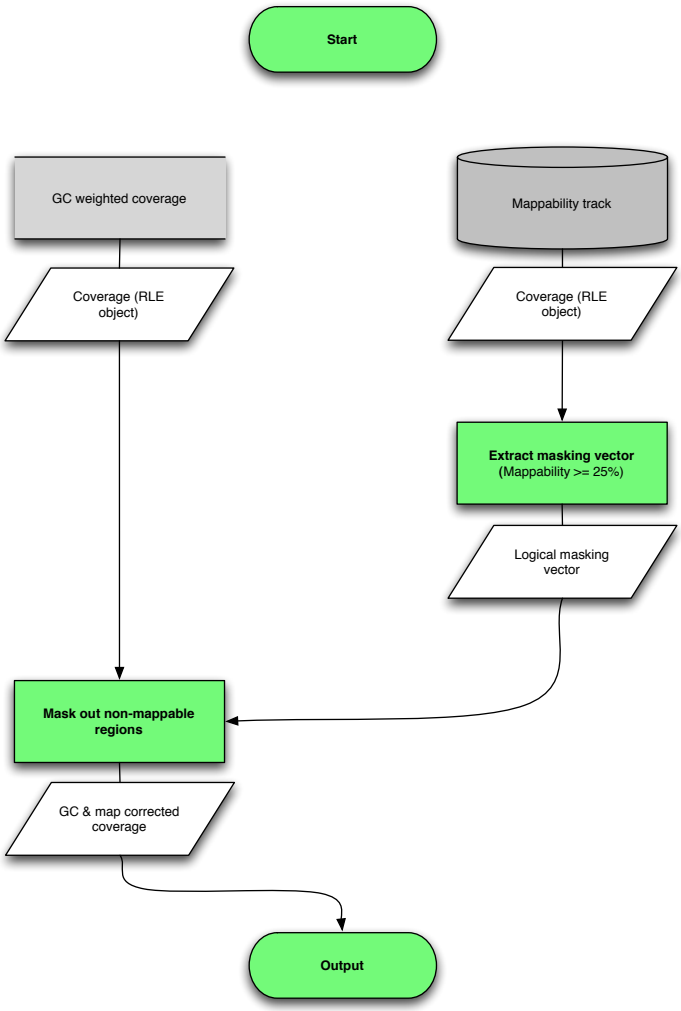


Figure 87 Non-mappable regions masking procedure shown as flowchart

The mappability correction step (**Figure 87**) takes as input GC corrected coverage and mappability track. The mappability track is used to build logical masking vector – the regions of mappability lower than 25% (mappability track coverage <100, where maximum mappability equals 400) are masked out. This masking vector is applied to GC-corrected coverage producing range length encoded list structure with masked out non-mappable regions.

4.3.7 The division step

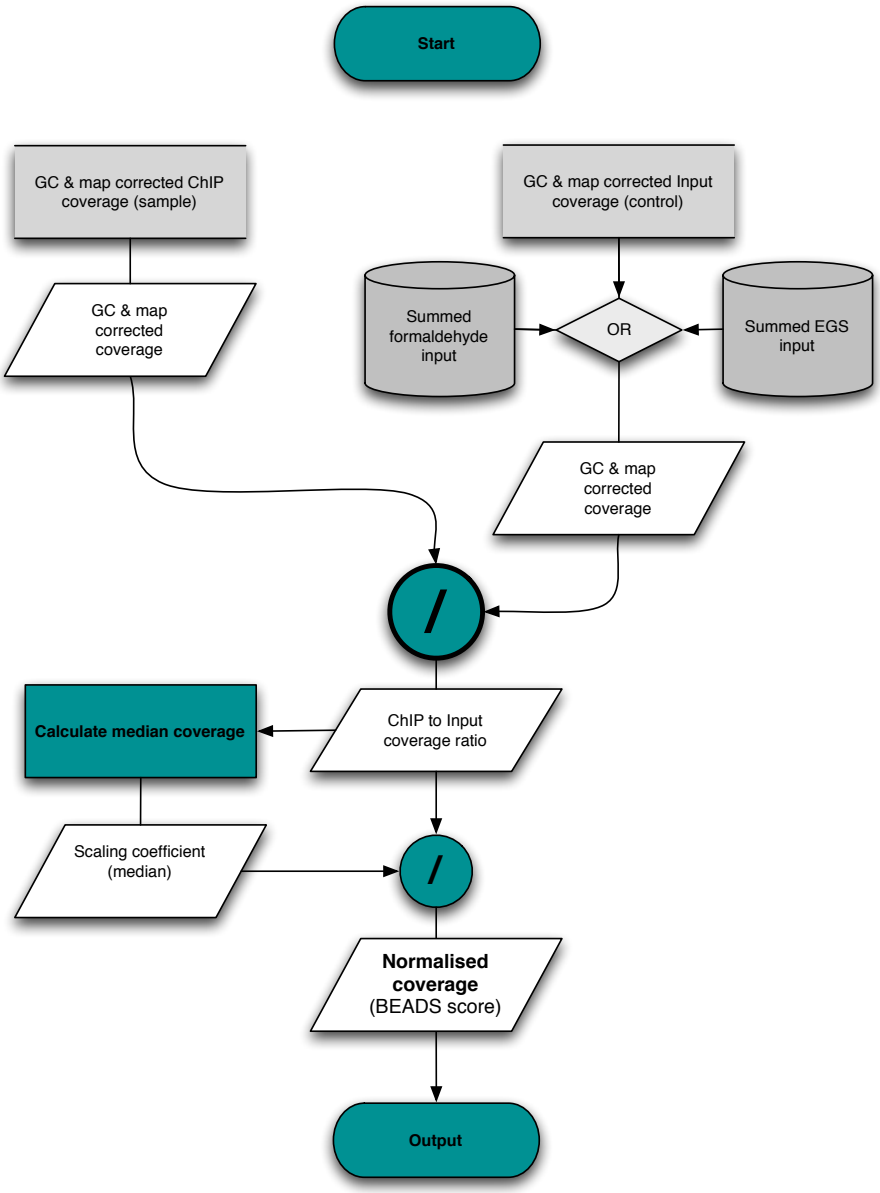


Figure 88 rBEADS division step shown as flowchart.

The final step of rBEADS implementation involves the division of ChIP experiment reads coverage track by experiment specific or summed *Input* coverage track (**Figure 88**). The ChIP to Input coverage ratio is produced. Both tracks are GC bias and mappability corrected using the previously described methods. Subsequently the median of ChIP to Input coverage ratio is calculated, providing a scaling coefficient. Then, ChIP to Input ratio track is divided by this value, producing the final BEADS score. This procedure scales the median ratio value to 1 providing meaningful enrichment

score in peak regions to determine enrichment/depletion in relation to the median value. Furthermore, using BEADS score enables to directly compare the tracks produced with different depth of sequencing datasets.

4.4 JADB – integrated database, data processing and visualization system

Genomics analyses are facilitated by collecting and analysing huge datasets. For ChIP-seq and RNA-seq we typically sequence 20 million reads. When calculating a coverage track this translates to a vector of more than 100 million numeric values for *C. elegans* and roughly 3 billion for *H. sapiens*. The ability to quickly, efficiently and automatically process this data is vital for effective using genomics resources – having as much as possible of quality check (QC), processing and summarising data allow to spend more time and resources to actually analyse them and use them as a resource to discover biological phenomena. Even more than processing, effective storage and retrieval of the data make them much more useful, and drastically increase speed and ease of analyses – the ability to obtain signal at genomic loci of interest, from pre-calculated track that was processed in the way best suited for given analyses, is a game changer in genomics. For example, after running ChIP-seq analyses I am storing 18 data entries per experiment and 11 track files that are compressed vectors of >100 million data points. This adds big overhead for storage and cataloguing requirements of the system but allows to find a track produced with set of processing parameters (such as different rBEADS normalisation options, and different scaling of data), which will be best suited for answering a certain biological question. Last, but not least, sequencing data are only as good as metadata attached to them. Without proper information what the data represents, what was the experimental design and what data processing was applied even the highest quality sequencing experiment is not useful for integrated analyses. This is the problem of many big sequencing consortia and data access

initiatives, like The Cancer Genome Atlas (TCGA), which delivers huge collection of high quality, uniformly processed genomic data, but patient metadata are often incomplete or inconsistent. In controlled laboratory environment, working on model organisms, where pool of possible metadata is considerably smaller this issue is easier to manage, but still is a far greater challenge than processing or storing data due to unpredictable human factor – some metadata data have to be inputted manually by experimenters, and input and validation methods offer enough flexibility to handle new user cases.

When engineering JADB system I aimed to address all three challenges mentioned above and I think I vastly succeeded. JADB system connects online spreadsheet based experiment metadata submission, automated processing pipeline, integrated with Illumina BaseSpace cloud service, efficient and secure data and metadata storage, user authentication, web browser GUI for searching and downloading data and application programming interface (API), as well as integration with many other tools like SeqPlots, genome browsers (IGV and JBrowse and Biodaliance) and other data analysis and visualization tools.

4.4.1 In-house data collection and JADB database system

I have built a database system (JADB) that allows automatic processing (alignment and normalization) of raw data, as well as programmatic access to experiments and their annotations. I consider this system to be a foundation for all of my and many of my colleagues' research projects, as efficient retrieval of data is vital for large-scale analyses. As of April 2018, I have processed and stored ChIP-seq experiments on >36 histone modifications and variants, >90 chromatin associated factors in different developmental stages and mutant strains, in addition to 461 RNA-seq expression profiles: a total of 2301 sequencing experiments, including inputs and controls. The

total of 38,635 individual files are registered in database system. The JADB system provides web-based GUI and supports data sharing and collaboration between multiple users. JADB can restrict user access or experiment editing permissions with an internal user authentication database.

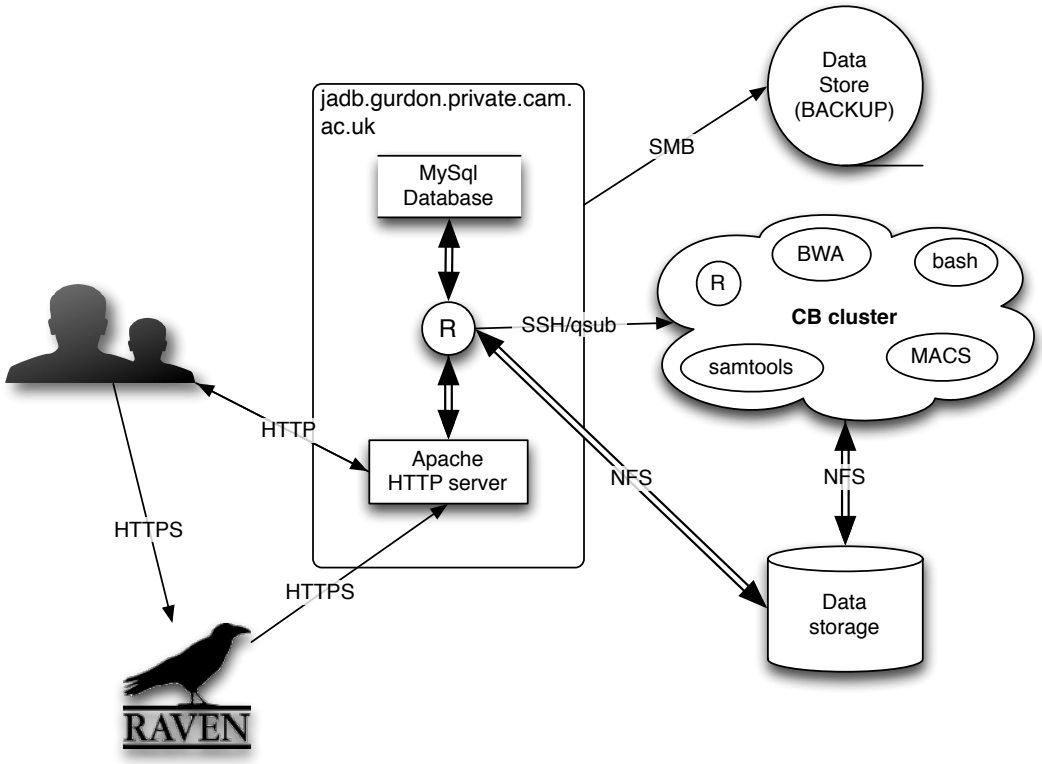


Figure 89 The ideogram of JADB database system and ChIP-seq/RNA-seq processing pipeline. The final users communicate with HTTP server (Apache) running rApache module for communication with R processes and MySQL database. R access network attached storage (NAS) for retrieving files and saving the results of analyses. The tools requiring heavy usage of computational resources are running on separate computational cluster. Finally, the data are periodically backed up using external storage service. This design has proven to be extremely robust, because it encapsulates frontend server, file storage and computational modules on 3 separate machines.

Moreover, it also allows the use of external authentication services, such as Raven ID authentication system provided by Cambridge University. This design of database offers great benefit to all members of our research group and external collaborators – it allows to easily share data with colleagues, get the feedback on data quality and usefulness from other members and re-purpose previous experiments for new projects. JADB has

easy to use, filterable and searchable interface allowing instant access to data, quality reports and results produced by integrated tools (**Figure 89**).

4.4.2 JADB implementation and data processing pipeline

JADB pipeline starts with user submitting metadata of ChIP-seq or RNA-seq experiments to JADB system. This can be done by providing a link to Google spreadsheet or such spreadsheet can be created directly in JADB interface. The metadata contain experimental information, and project/adapter ID for Illumnia Basespace cloud service. The entry is validated for metadata consistency and with Basespace that the indicated run exists. If validation is successful raw FASTQ files are downloaded from Basespace and registered in JADB. At this step an experiment entry is created with relational SQL database. This entry is linked with Extract and Collection tables, or relevant entries are created based on provided metadata if given extract/collection was not registered to JADB before. The Experiment table holds a unique experiment identifier and all metadata specific to experiment. Experiment entry serves as container for files that are inputs, intermediates or outputs for analyses. First file entry created and linked to Experiment entry is a FASTQ file. Instead of using Basespace, raw files can be also added from gene expression omnibus (GEO) or directly uploaded by users (**Figure 90**).

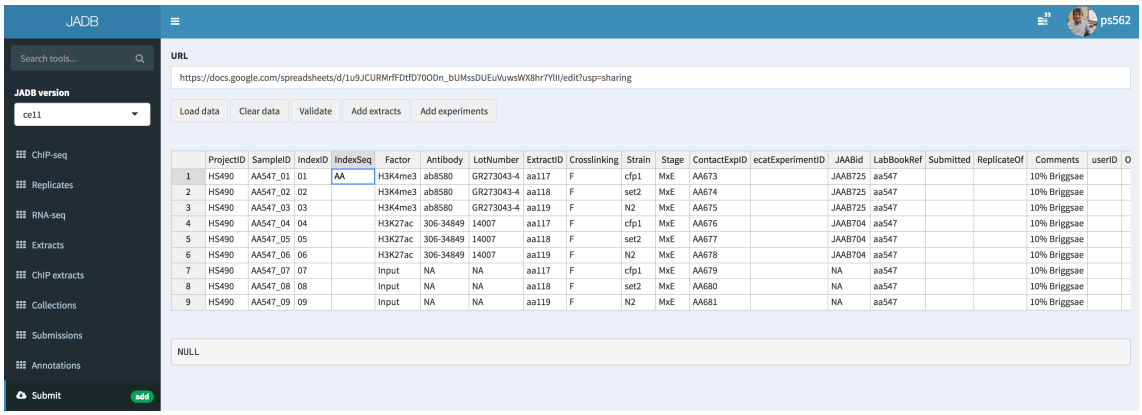


Figure 90 JADB graphical user interface for submitting new experiments

Files are stored physically on the hard drive of the machine running the JADB server instance, or at network attached storage (NAS) mounted on JADB server. They are stored in specially designed folder structure, which enables intuitive manual navigation and automated querying even without using SQL database – this is especially useful for lightweight shell scripts that can perform maintenance or data retrieval tasks. When raw input file is registered, default analyses pipeline begins. Each output file is registered to JADB and linked to experiment as pipeline progresses.

For ChIP-seq experiments the pipeline begins with alignment of FASTQ files to the selected reference genome, creating indexed binary sequence alignment/map (BAM) file. Then reads are extended to 200bp (expected insert size) and two coverage tracks are calculated: (1) “Q10”, which filters reads below mapping score of 10, efficiently removing all multi-mapping and lower quality reads, and (2) “NQ” that retains all aligned reads, which is a necessity when analysing repetitive regions of genome. Then, rBEADS pipeline is run three times with different settings – mentioned before “NQ” and “Q10”, plus another run of Q10 with the additional step of uniquing the reads, i.e. allowing only a single read mapping to give base position, on given strand in given chromosome. This can be useful to remove PCR duplicates. However, it caps the signal at 400 reads maximum (2 times 200bp for each strand), hence it can artificially reduce and flatten signal peaks in very narrowly positioned factors, like polymerase II, where the same read might have been independently generated and not be a PCR duplicate. Each run of rBEADS produce BEADS score signal tracks (linear scale), log2 scaled tracks, and z-scored tracks – a total 9 tracks with different normalisation and scaling settings. The next step is running MACS2 peaks calls software that produces peak calls in narrowPeak format and peak summits in BED format. Finally, sequences from equal regions of 500bp around peak calls are collected and fed as input to MEME-ChIP de-novo and database associated motif calling pipeline. Parallel to data processing, two QC

utilities are run – FastQC to assess general quality of sequencing run and fastqSCREEN, that aligns representative sample of reads to multiple genomes, including bacteria, fungi, mouse, human, artificial sequences etc. to detect possible contamination of samples.

For RNA-seq data raw FASTQ are filtered from adapters using Trim Galore. Then, trimmed FASTQ is aligned with STAR aligner – this step produces an indexed BAM file, which is subsequently used as input for tag counting algorithms and later differential expression analyses. Tag counts for coding genes annotation are saved as CSV table. When required for analyses, other annotations can be used with the DEview tool – a GUI application for correlation analyses and visualization. Also, the read coverage tracks are created and deposited in JADB as BigWiggle files. For experiments focusing on mRNA reads pileups and reads per million normalised reads pileups are created. For strand specific libraries also stranded tracks can be produced. These tracks can be further visualised with SeqPlots or genome browsers. Specialised tracks are created for unusual types of libraries, such as those mapping transcription initiation, where tracks representing the first base pair of the read are needed.

JADB works great with high performance computing (HPC) resources – it uses Slurm job queuing system to run many jobs in parallel on different nodes. The recommended setup is to run JADB server on dedicated virtual machine or inside a Docker container and utilise NAS for storage and cluster setup for computations. However, the whole system can also run on single desktop computer, as long as its hardware configuration can handle data processing pipelines.

4.4.3 End user features and functions of JADB

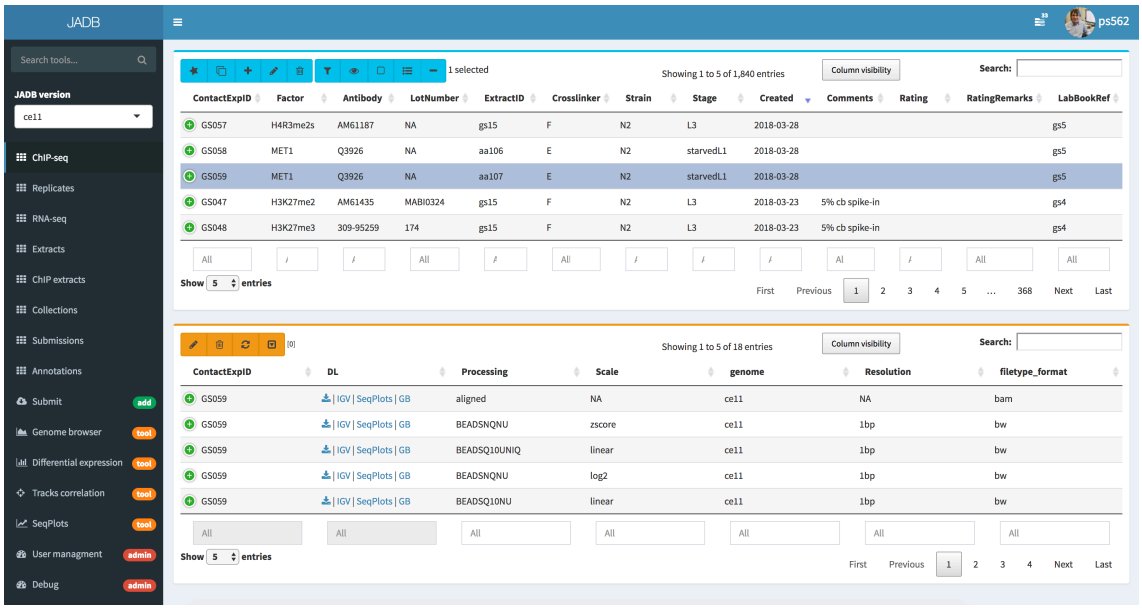


Figure 91 JADB user interface as seen in the web browser. Top bar shows user information, and tasks performed by server, like data processing jobs. Toolbar on left allows to swhich between genome versions, types of experiments and engage the visualization and analyses tools. In the main view blue panel (top) shows the experiments, ChIP-seq in this instance – experiment highlighted in light blue is selected. Orange panel (bottom) displays files associated with the selected experiment.

The main benefit of JADB is providing a lab centric repository for all experiment types, and annotations. JADB system supports multiple genomes – in our lab I am using it to support two different genome assemblies of the *C. elegans* genome, with full data processing pipeline and storage done independently for these versions. It can be also used for storing data from multiple organisms. It uses Google Spreadsheets as metadata and experiment submission system, while saving all data locally. Alternatively, users can input metadata directly in the system, and use internal submission user interface. JADB also features a rating system, allowing users to flag potentially failed or problematic experiments without a need to alter core experiments metadata, i.e. having permission to edit given experiment. This way we can crowd source system users to help find problems and keep good quality of data - this is very useful when managing larger pools of experiments, where automated QC cannot spot all problems.

JADB uses web browser interface (**Figure 91**), and has a full feature application programming interface, which makes it efficient to integrate with external tools. For example, I have integrated the Biodaliance genome browser (Down et al 2011) within JADB web interface, which allows quick inspection of genomic data. I also provide IGV integration, which loads the data directly to IGV session running on user's personal computer, without a need to download any data. I also developed and integrated into JADB user interface visualization and data analyses tools: (1) SeqPlots, which can load data directly from JADB, (2) CorTracks for correlation analyses and (3) DEview for RNA-seq differential expression time series analyses. JADB can run within Docker container, which makes it easy to run the whole system on any server configuration, or locally on a desktop machine.

4.4.4 Automated replicates detection (ARD) method

The database has a high number of experiments, that are testing for the same factor in same strain and growth condition but were performed using different antibodies or with modified protocols. Many experiments have at least two biological replicates which need to be combined. In data analyses techinques, especially automated ones relying on machine learning, it is extremely important to use high quality data, i.e. the experiments that represent the biological data most accurately. However, manually selecting best replicates, and then best antibodies and laboratory conditions for given factor is very laborious and hard to keep consistent in large datasets. For these reasons I decided to engineer and implement an automated method for detecting and merging best replicates in JADB.

In the first instance the method scans through all ChIP-seq experiments in JADB and selects those where there are at least two independent experiments matching following criteria: (1) same factor is immunoprecipitated, (2) same antibody is used, (3)

experiments were done in same strain and (4) same developmental stage. They are grouped together, and in these groups for every possible pair I calculate following statistics: (a) Pearson correlation coefficient - r, (b) replicates peaks overlap coefficient - p, (c) sum of peaks in both replicates - n, (d) and summary coefficient. Replicates peaks overlap coefficient is calculated as ratio between peak intersection (genomic regions covered by peaks in both replicates) and peaks union (genomic regions covered by any replicate) and takes value from 0 to 1, where 1 denote perfect overlap and 0 no overlap at all. The summary coefficient is calculated using following equation:

$$coef = r^2 + p^2 + \left(\frac{n}{n_{max}}\right)^2$$

Equation 1 Summary coefficient for automated replicates matching algorithm. n_{max} represents highest value of n within the replicate candidates group.

Next, replicates with r value below 0.75 are discarded. From these experiments, the replicate pair with highest value of $coef$ is selected as best representation of replicated signal in the group. Groups with fewer than two well correlated replicates are discarded. Also, groups with mismatching experimental protocols are discarded. Then, for all matching experiments a pipeline to combine replicates is run, and replicates are registered in JADB. This still leaves replicates representing the same stage, strain and factor combination, and immunoprecipitated using different antibodies (for example, H3K27me3 in L3 stage and N2 strain has 4 good replicates in HK00013, UP07-449, 1E7 and AB6002 antibodies) but it is very easy to use summary coefficient to select antibody producing most matched replicates, or use antibody validation information to select one manually.

4.5 Correlation analyses

Pearson correlation coefficient, ranked coefficients - Spearman's rank correlation coefficient and Kendall rank correlation coefficient, maximal information coefficient (MIC) (Reshef *et al.* 2011) and distance correlation (dcor) (Székely *et al.* 2007) are common measures to assess the similarity of two vectors. In general, these methods produce values ranging from -1 to 1 on a continuous scale, where 1 indicates perfect similarity, -1 - perfect dissimilarity and 0 - no association between vectors. The final output is similar, but the way it is calculated vastly differs between different methods, which is reflected on their computational requirements. The most efficient is Pearson correlation coefficient, which implementations relies on simple matrix algebra, is computing power and memory efficient, scales great with data size, and thus can be used easily for high dimensionality datasets. Ranked correlation coefficient are moderately memory and CPU efficient, as they require to sort the vector, and scales roughly linearly with vector size. On the opposite spectrum are MIC and dcor require complex computations (e.g. MIC relies on information entropy estimation) and are not viable for very high dimensionality.

Since ChIP-seq and RNA-seq tracks can be represented as vectors, and set of experiments as matrices, these methods have a straightforward application for sequencing experiments data analyses. I already discussed the usage of Pearson correlation coefficient as a good metric for replicate similarity. Extending these pairwise analyses to bigger collection of data gives us a correlation matrix as output, which can be interpreted directly, giving us some information of global data structure, e.g. associations between experiments, plotted for visualization, or passed to more advanced machine learning algorithms for further processing. The advantage of correlation matrixes is great dimensionality reduction. For example, in *C. elegans* 100 ChIP-seq experiments at 1bp resolution are represented as 100x100,000,000 numeric matrix,

summing to 10e+10 data points. After calculating correlation, we obtain 100x100, symmetric matrix, in which all diagonal elements equal 1, meaning that we efficiently reduced previous high dimensionality dataset to just 4950 correlation coefficients, giving impressive 4.95e+7 times reduction in data size. The trade-off for this is that we lost all local interactions in the genome – we have only a single number describing the similarity between all experiments in our cohort. In further chapters I will focus on methods allowing an impressive dimensionality reduction without losing any information about local interactions. Nevertheless, correlation analyses are still a very useful technique, because of simplicity of output and intuitive understanding of results. Indeed, in some cases we might greatly benefit from ignoring local similarities and focusing on global structure of data.

4.5.1 Data retrieval and summarization techniques

Efficient retrieval of ChIP-seq profiles and RNA-seq data in vectorised form is an important pre-requisite for further statistical processing. Also, binning of signal might be vital for successful implementation of many algorithms because computational limitations cannot accept genomic data at single base resolution. Assuming such resolution, single ChIP-seq experiments produce around 100M data points for *C. elegans* and more than 3G data points for mammalian organisms. Using such high dimensional vectors and arrays as the input might be variable for simple aliases, such as calculating Pearson correlation coefficients, but using more sophisticated machine learning techniques, such as MIC or Bayesian machine learning algorithms, will be computationally challenging, even assuming we can parallelize the algorithm efficiently, which is often not possible.

Furthermore, the ChIP-seq and RNA-seq data tend to be noisy at a single base level and implementing a data summarization technique can alleviate this problem. Smoothing the

data, hence removing a single base pair resolution noise will actually increase the power of most machine learning techniques to generalize on input data.

The simplest way to address this problem is averaging the profile with reasonable size bin, i.e. the size that best captures the width of the feature of interest. For example, to capture a chromosome modification the distance ~150bp (the length of DNA around a nucleosome) might be the best choice.

Conveniently, the ability to rapidly extract signal values at a given resolution in given genomic loci is already implemented in BigWiggle (BigWig) track data format. BigWig enables data acquisition using a multi-layer summarization, which allows for significant speedup while binning the ChIP-seq profiles and RNA-seq coverage tracks (Kent *et al.* 2010). BigWiggle is a member of Big Binary Indexed (BBI) files family – compressed, binary, indexed files that contain the data at several resolutions. BigWig and BigBed implements multi-layered summarization approach, combined with various indexing methods (primarily B+ trees and R trees), caching (sparse files) and compression (zlib). They also store chromosome size information, which allows checking if files were produced using the same reference genome version. The summarizations include windows of 100 bases, 200 bases at a time, etc. and stores summaries like minimum and maximum signal value, sum of values, number of bases covered with signal. These make it extremely efficient to calculate statistics like mean standard deviation, coverage, etc. Similarly, binary alignment maps (BAM) format for storing alignments can be efficiently queried for retracting reads at given genomic loci and on-fly coverage calculation.

I designed JADB to extend the advantages of BBI storage from single file to collection of experiments. JADBtools R package provides API to JADB. Instead of retrieving a vector representing summarised experiment, JADB provide interface to retrieve a

matrix, representing a collection of summarised experiments, and contains experiment annotation retrieved from SQL database. As the result of combining the BigWiggle file infrastructure, with efficient storage and retrieval, the system is extremely fast – I benchmark it on my laptop (using high speed Wi-Fi and 6 parallel threads) achieving speed of querying up to 50MB/s allowing to retrieve 92 experiments at 100bp resolution (1002861x92 matrix) in less than 5 minutes.

In addition, this package allows searching and retrieving whole experiments and analyses results from centralised storage via direct HTTP. Further, it can retrieve numeric vectors, matrixes or GenomicRanges structures from files using I/O library for BBI, which allows higher flexibility and adjusts retrieved data to requirements of further analyses. In addition, JADBtools provides some high-level analyses tools, including QC and reporting tools, motif analyses, differential expression analyses run directly on RNA-seq experiments in JADB using DESeq2 and edgeR, and peaks calling using external software and peak replicates summarization with intersection or IDR approaches.

4.5.2 Binnig size in genomics data has an impact on correlation analyses

As discussed in previous chapter, binning is a straightforward technique, that allows us to remove single base resolution level noise and decrease dimensionality of input datasets, hence efficiently reduce computational workload and memory footprint of machine learning algorithms. However, the size of binning window has to be well selected to represent the nature of the data and relationship we want to capture and analyse.

To establish what is the effect of different binning resolutions I conducted a simple experiment – I have selected similar, well correlated (HPL2 and LIN61), anti-correlated

(H3K27me3 and H3K79me2) and not correlated (H3K9me2 and Pol II Ser2P) tracks from replicates database, binned them at different resolutions and calculated Pearson correlation coefficient (**Table 21**).

Bin size	HPL2 & LIN61	H3K27me3 & H3K79me2	H3K9me2 & Pol II Ser2P
no binning	0.91423	-0.30242	-0.00032
10	0.91479	-0.30490	-0.00045
100	0.92116	-0.32979	-0.00074
1000	0.95647	-0.49395	0.00012
10000	0.97525	-0.67447	0.02368
100000	0.98467	-0.69686	0.06631

Table 21 Impact of binning resolutions on Pearson correlation coefficient. The columns show well correlated (HPL2 and LIN61), anti-correlated (H3K27me3 and H3K79me2) and not correlated (H3K9me2 and Pol II Ser2P) tracks from replicates database.

In case of similar tracks, binning did not drastically affect the coefficients. However, in case of anti-correlated tracks, i.e. H3K27me3 – repressive mark and H3K79me2 – marking active genes, correlating using high binning size better captured the negative correlation. This is because large domains are anti-correlated, while local noise signal seems not to be correlated at all, or even weakly correlated on smaller scale due to similar noise profile. Not correlated tracks, like H3K9me2 and Pol II Ser2P were least affected by selection of the bin size – their correlation remained close to 0 in bin ranges from 1bp (no binning) to 100kb (**Table 21**).

Binned correlation can be also used to validate replicates. Similarly to well correlated replicates tracks, the selection of binning size have no detrimental impact on calculated correlation (**Figure 92**). 1kb or 100bp are good choices, as they provide good trade-off between resolution and computational complexity.

As shown in the examples, the binning size selection has significant impact on output of statistical methods. The appropriate binning size should take into account the resolution

of the analyses, the data we are representing, and the computational complexity. Also, different types of analyses might require different binning sizes. When simulating complex associations between the tracks, it might be best to use an adaptive model that will infer the optimal binning size from data or hold sets of values calculated in different bins. Also, Bayesian machine learning models, which infer latent structure of the data and incorporate noise models are much less vulnerable to noise driven data correlation at high resolutions.

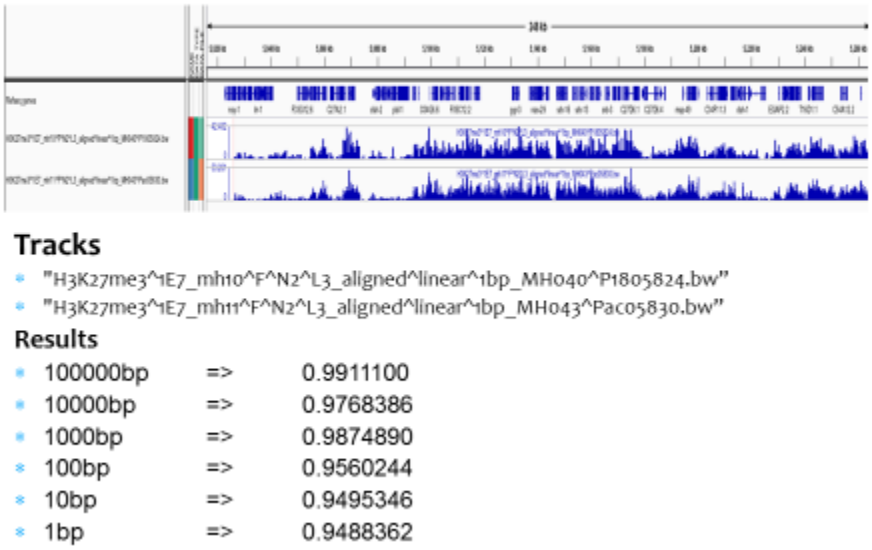


Figure 92 Replicates correlate well regardless of bin size. The correlation is higher at lower resolution, but even at single base the difference is minor. Values represent Pearson product-moment correlation coefficient.

4.5.3 Global PCA analyses for ChIP data reveals the structure of basic associations

In contrast to many existing approaches that analyse genomic regions, *e.g.* chromatin state analyses, which focus on clustering the genome based on chromatin marks, I focused on finding profiles exhibiting similar patterns, therefore being likely to share related functions. Using data acquisition methods characterised above I integrated continuous signal profiles in single, high dimensional data matrices: histone modification and TF binding tracks, with categorical data, e.g. genomic feature. I decided to use ChIP-seq profiles as continuous signal, rather than as binary data

produced by running the peak calls. rBEADS normalised ChIP-seq signal should be a good proxy to the relative level of occupancy of transcription factors and histone modifications.

Before starting with further analyses, I examined the data using principal component analyses. I acquired a 1kb resolution matrix for all unique factors (61) from third larval stage (L3) wild type (N2) worms in JADB from z-scored tracks. This produced 100269x61 data matrix. Then I computed the singular-value decomposition of input matrix, obtaining matrix **v** whose columns contained the right singular vectors, or components (PC) and matrix vector **d** containing the singular values. The two first columns of v were used as dimensions of PCA plot, and **d** was used to calculate percentage of variance captured by components, according to following formulas (where $\boldsymbol{fv}$ denoted fraction of variance and $\boldsymbol{i}$ the number of component):

$$fv_i = \frac{{d_i}^2}{\sum d^2}$$

I observed that the PCA plot shows some underlying structure of the data. To extract this structure, I used k-means clustering. I determined that the best and most interpretable results were achieved using the first three components for clustering (PC1 – PC3) and five clusters. This de-novo analyses revealed interesting separation of the ChIP-data (**Figure 93**). Cluster 1 (red) is composed mostly of histone marks and factors deposited on gene bodies or on extended Pol II elongation regions, such as H3K36me and H3K79me. It also contains Pol II Ser5P ChIP – a specific antibody recognising RNA polymerase II C-terminal domain (CTD) serine 5 phosphorylation – a covalent modification of Pol II characteristic of polymerase initiation and early transcription cycle.

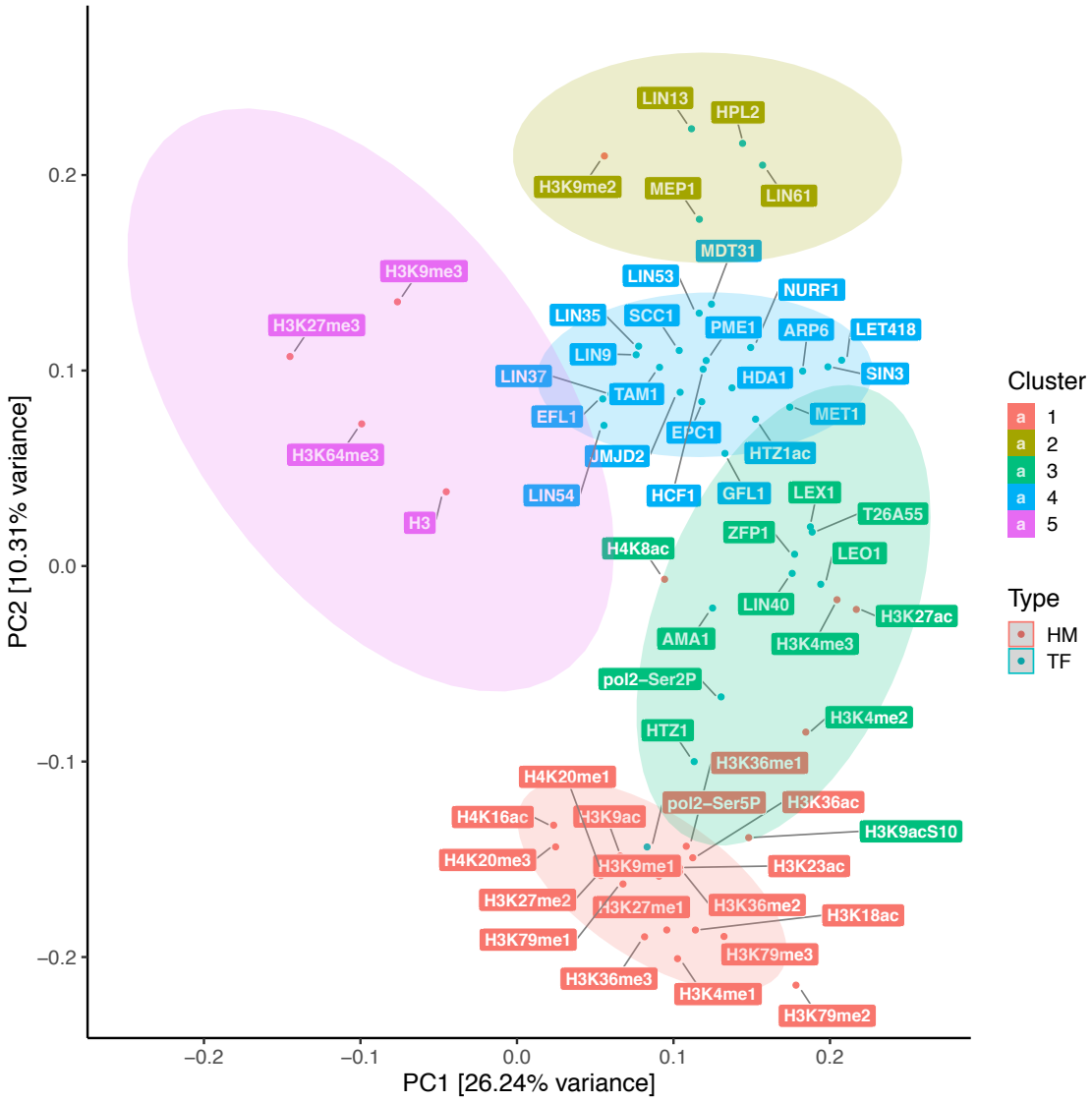


Figure 93 PCA plot for WT, L3 larva ChIP-seq profiles acquired at 1kb resolution. Principal components 1 and 2 are plotted, and the tracks are closeted with k-means into 5 clusters using first 4 components. Clusters are represented as colour ellipses, and data assignment is colour coded. Ellipses are 2-dimensional projections of 3-dimensional clustering space (PC1 - PC3), so they might overlap each other.

Cluster 2 (yellow) is composed of heterochromatin factors that were discussed in detail in chapter 2 – HPL-2, LIN-61, LIN-13, H3K9me2 and MEP-1 – a component of the Mec complex. Cluster 3 (light green) is composed of chromatin factors and chromatin marks found at transcription initiation and promoter regions. Here we can find marks prevalent on TSSs, such as H3K4me3, H3K27ac, HTZ-1, chromatin regulators such as LEX-1, ZFP-1, and LEO-1 and RNA polymerase itself – both immunoprecipitating against globular domain AMA-1 and against Pol II Ser2P - C-terminal domain serine 2

phosphorylation, which is a hallmark of productive transcription elongation and shows highest abundance toward the 3′ ends of genes. Cluster 4 (sky blue) gathers various chromatin and transcription factors and a histone variant modification – HTZ-1ac (H2AZ). These include members of DREAM complex, MET-1, HDA-1, SIN-3, HCF-1 discussed in chapters 3 and 4, and LET-418 discussed in chapter 2. Finally cluster 5 (purple) contains heterochromatin marks – H3K9me3, H3K27me3, H3K64me3 (Lange *et al.* 2013) and histone H3, which indicates histone dense regions.

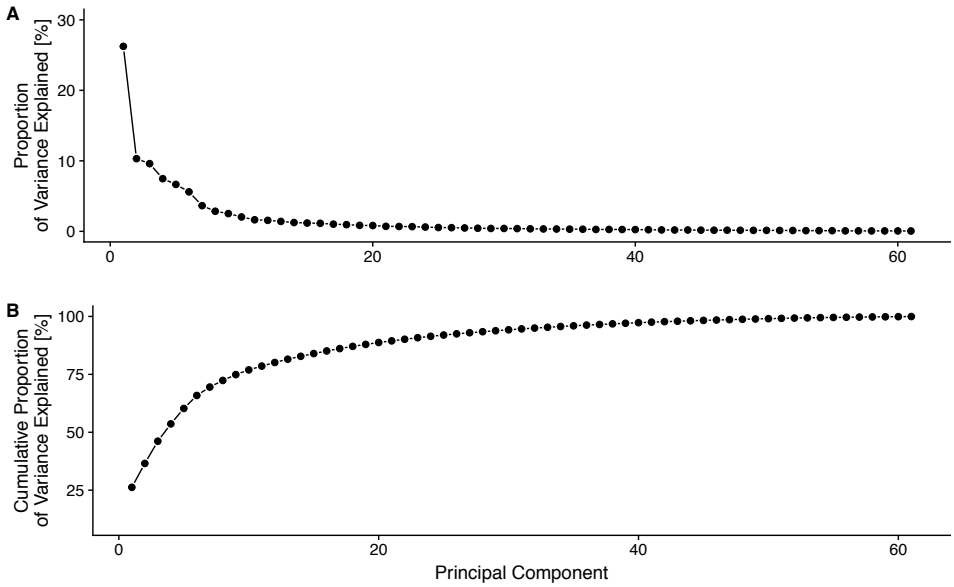


Figure 94 Diagnostic plots for PCA acquired with 1kb binning resolution. (A) Proportion of variance explained plot and (B) Scree plot showing cumulative proportion of variance captured by given number of principal components.

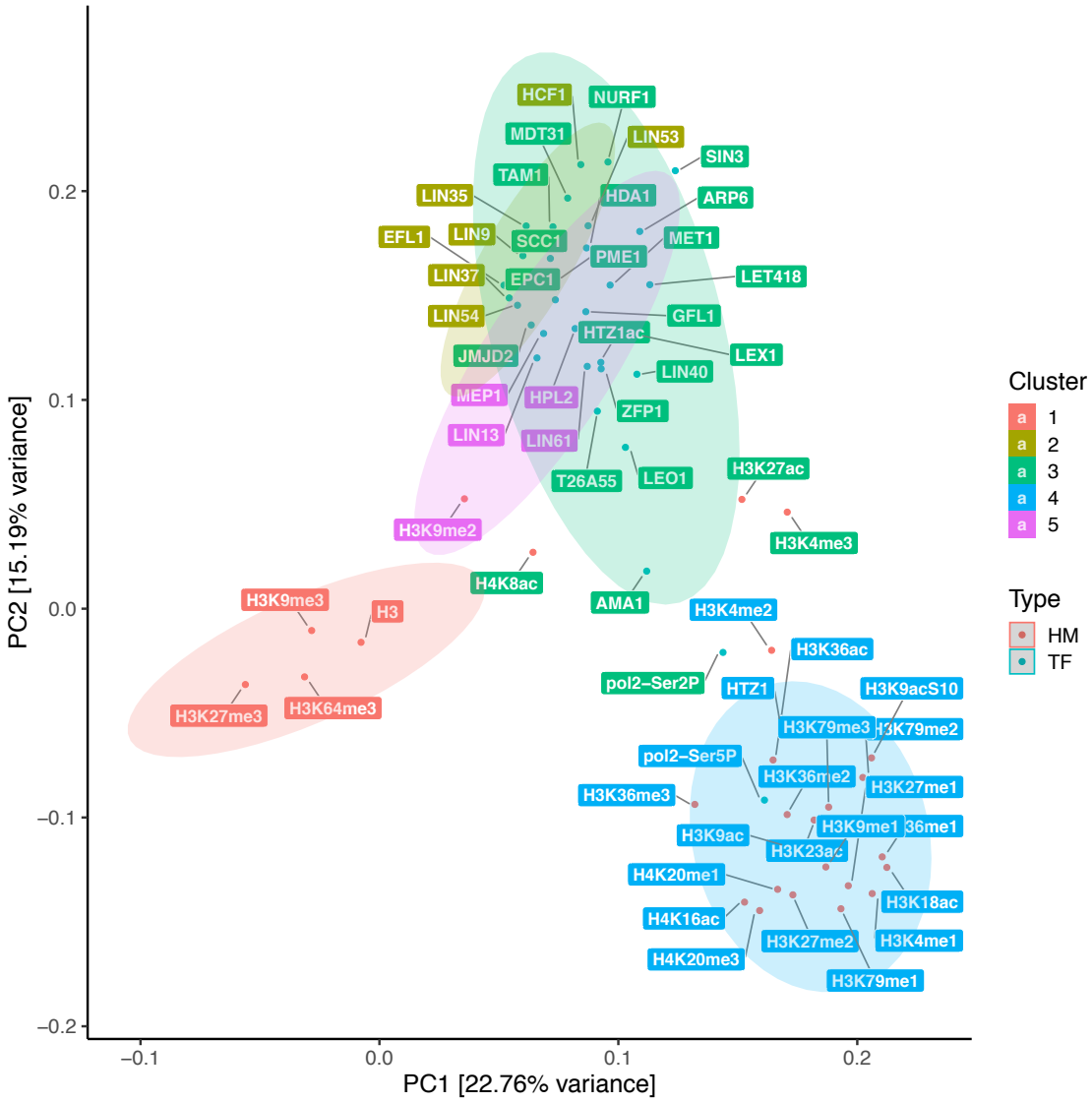


Figure 95 PCA plot for WT, L3 larva ChIP-seq profiles acquired at 100bp resolution. Principal components 1 and 2 are plotted, and the tracks are closeted with k-means into 5 clusters using first 4 components. Clusters are represented as colour ellipses, and data assignment is colour coded. Ellipses are 2-dimensional projections of 4-dimensional clustering space (PC1 - PC4), so they might overlap each other. The clustering is similar to previous plot, but first 4 PCs have to be used to capture enough variance for clustering.

To better assess the effect of binning window and assess interactions between tracks in higher resolution I repeated the PCA analyses with 100bp resolution producing 1002724x61 data matrix (**Figure 95**). On first inspection I observed, that on the PCA plot showing PC1 vs. PC2 the separation of the classes is not as good as on 1kb bins-based plot. Indeed, PCA diagnostics (**Figure 96**) showed that the first two component capture smaller percentage of variance than in 1kb bins plot (**Figure 94**). Consistent

with this observation, I was able to re-create very similar clustering as shown with 1kb bins, but I needed to use first four components instead of first three – the clusters have different labels, as they are assigned at random, but generally five clusters in 100bp bins analyses were composed of same classes of factors. This is reflected on PCA plot (**Figure 95**), where clusters 1, 4 and 5 overlap each other in two-dimensional projection but are still well separable in 4-dimensional principal component space.

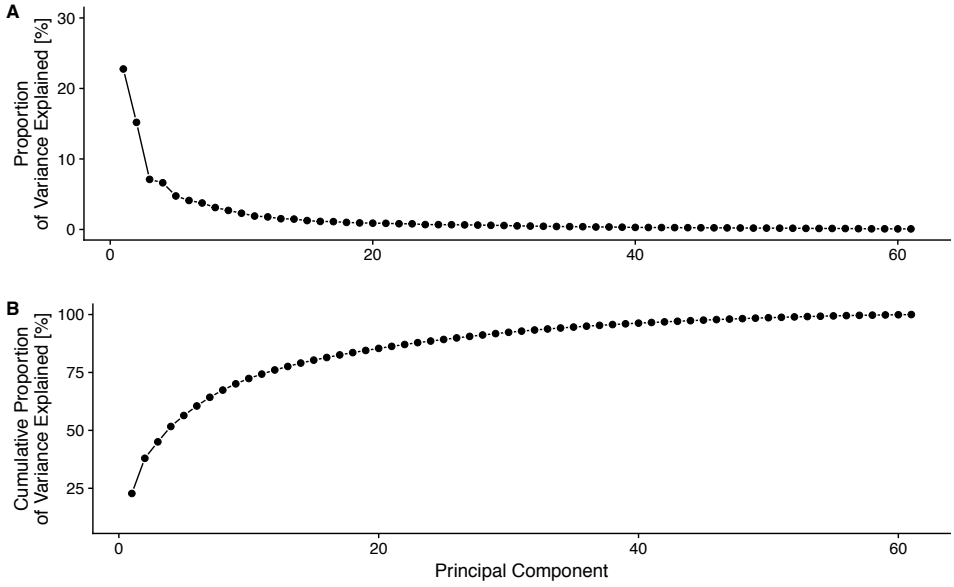


Figure 96 Diagnostic plots for PCA acquired with 100bp binning resolution. (A) Proportion of variance explained plot and (B) Scree plot showing cumulative proportion of variance captured by given number of principal components.

4.5.4 Global correlation analyses for ChIP data reveals structure

The PCA analyses shows structure within our ChIP data that can be retrieved using fairly simple clustering methods, i.e. k-means. However, these analyses rely on the analyst to choose parameters accordingly to data structure – what number of components and what number of clusters will be the best to describe data. Also, as we have seen with 100bp bins-based example, the 2-dimensional, or even 3-dimensional PCA plots can obstruct the true structure of more complex data, as the important variance might be hidden in higher components. For this reason, I wanted to move to non-parametric approaches, starting with correlation analyses.

The input for this method is the same matrix as for PCA analyses. I start with calculating Pearson product-moment correlation coefficient matrix and nonparametric correlation matrix using Spearman's rank-order correlation coefficient implemented in R. Further, I determined the statistical significance of each correlation using correlation test - a t-test is applied to the individual correlations using the following formula:

$$t = r * \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

This method is implemented in “psych” package in R. The resulting correlations and associated p-values are visualised for further inspection using “corrplot” package in R (Epskamp *et al.* 2012). The variation of plot I use is a heatmap-like plot, which shows correlation values in lower triangle and representation of correlation as circles in upper triangle, where size is proportional to absolute correlation value and colour to correlation strength, where blue denotes anti-correlated experiments, white – no correlation, and red – positive correlation. Upper triangle also shows statistical significance estimate – p-values are encoded as asterisks, where “*” denotes p-value lower than 0.1, “**” – 0.01 and “***” – 0.001.

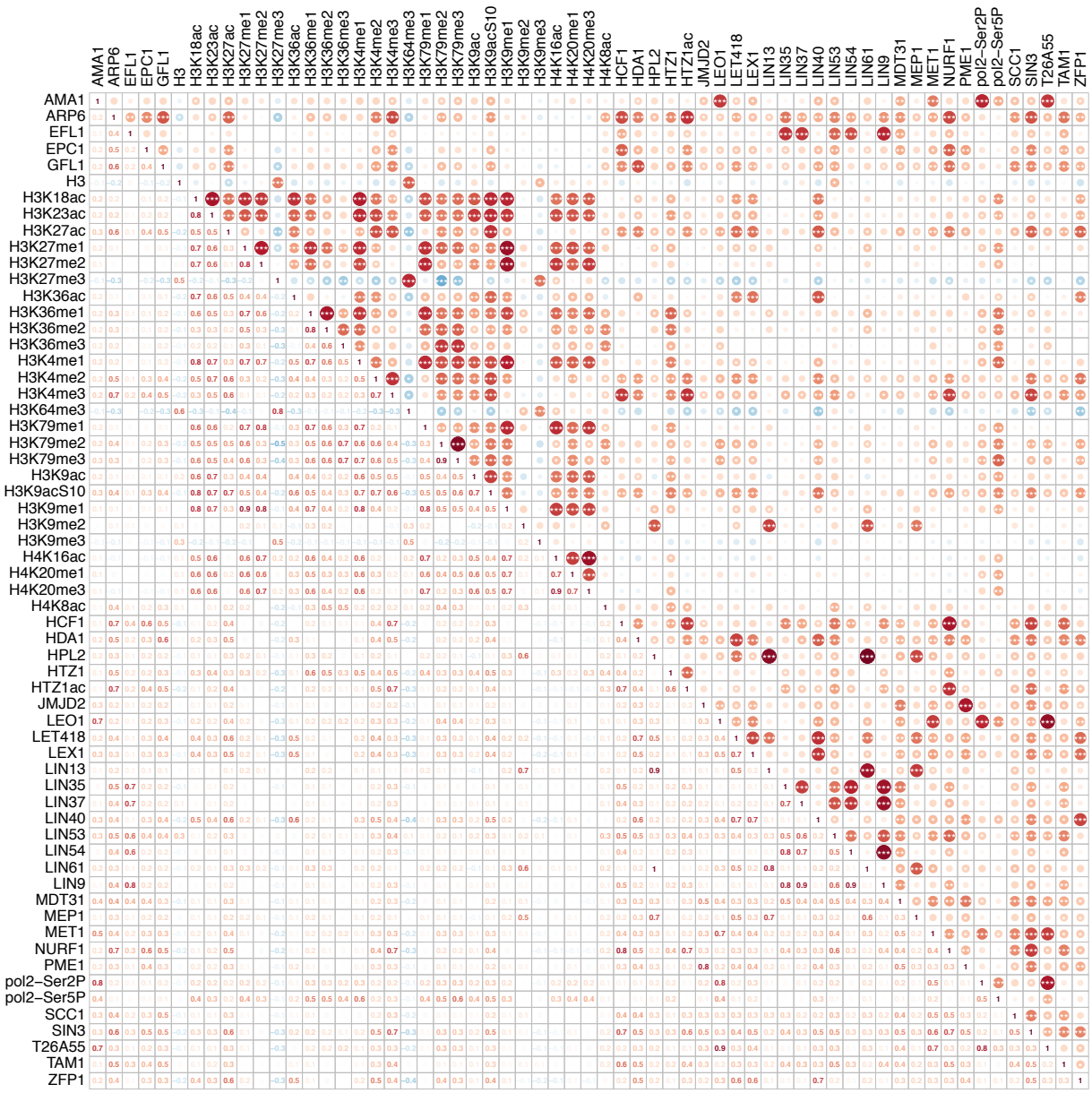


Figure 97 Correlation diagram at 1KB resolution for 61 factors immunoprecipitated in wild type in L3 larva stage. Correlation values are shown in the lower triangle and representation of correlation as circles is shown in upper triangle, where size is proportional to absolute correlation value and colour to correlation strength, where blue denotes anti-correlated experiments, white – no correlation, and red – positive correlation. Upper triangle also shows statistical significance estimate – p-values are encoded as asterisks, where "*" denotes p-value lower than 0.1, "**" – 0.01 and "***" – 0.001.

Similar to previous PCA analyses, I acquired signal at two resolutions, 1kb and 100bp bins. In accordance with PCA analyses we can clearly see a structure in ChIP profiles – there are groups of factors that correlate with each other and are all anti-correlated to other groups of factors (**Figure 97** and **Figure 98**). The comparison between 1kb bins

and 100bp bins shows an interesting trend – the 1kb bins are more likely to capture weaker association between groups (**Figure 97**), while 100bp bins show stronger correlation within the groups, and weaker between groups.

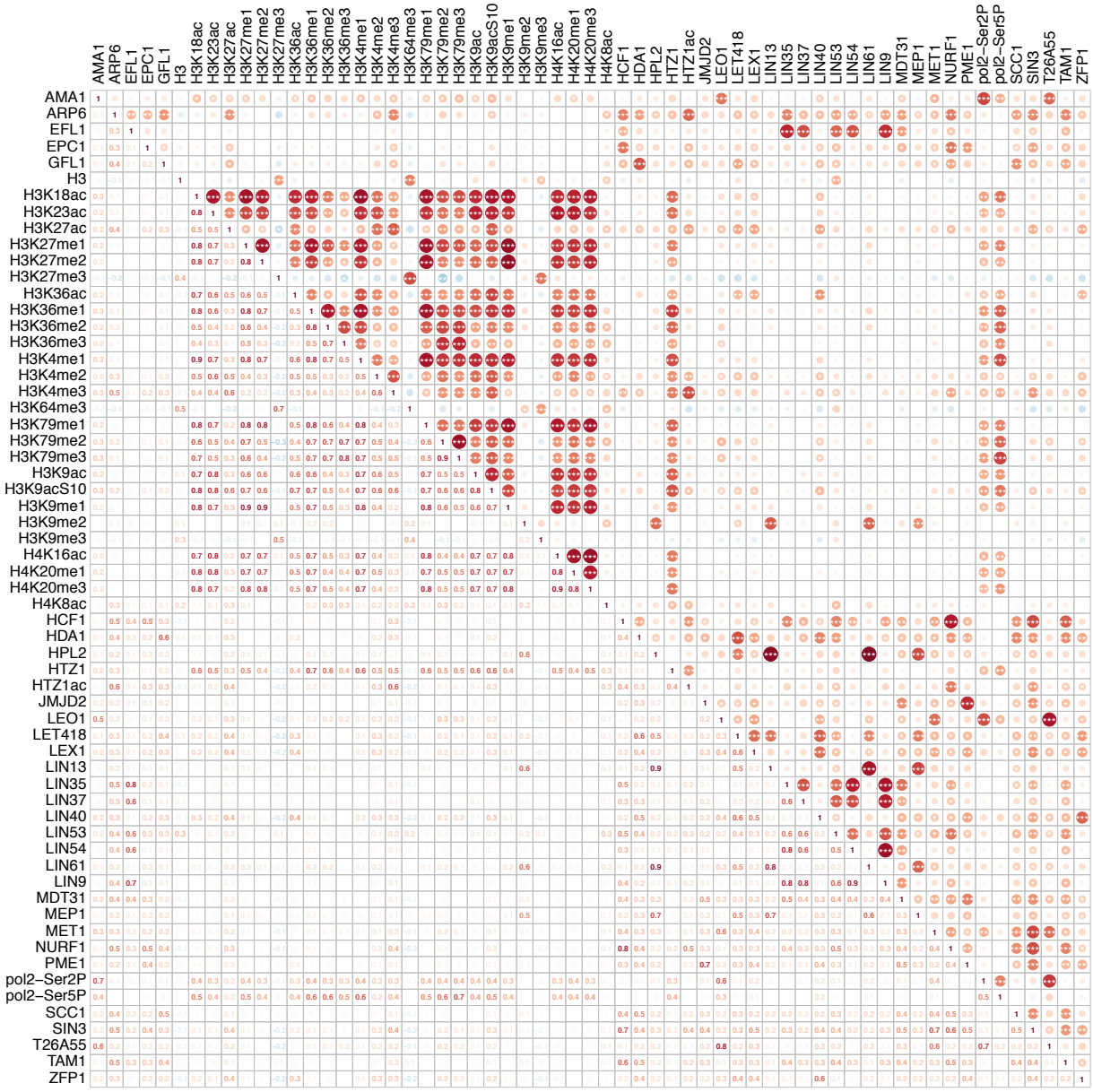


Figure 98 Correlation diagram at 100bp resolution for 61 factors immunoprecipitated in wild type in L3 larva stage. Correlation values are shown in the lower triangle and representation of correlation as circles is shown in upper triangle, where size is proportional to absolute correlation value and colour to correlation strength, where blue denotes anti-correlated experiments, white – no correlation, and red – positive correlation. Upper triangle also shows statistical significance estimate – p-values are encoded as asterisks, where "*" denotes p-value lower than 0.1, "**" – 0.01 and "***" – 0.001.

4.5.5 Clustered correlation analyses

In the next step I clustered the heatmap based on correlation and assayed if we can reproduce the groups found during PCA analyses and extract some more informative and detailed structure.

For this analysis I turned to hierarchical clustering (HC), which in contrast to k-means does not require the number of required clusters to be given *a priori*. Rather than assigning the data into clusters, HC organises data into a hierarchical structure using distance between data vectors. This structure is usually visualised using dendrograms and can be turned into different numbers of clusters after clustering is done with a cut tree algorithm.

In my clustering method I simply turned a correlation matrix to distance by applying $D = 1 - R$ transform, where D is a distance matrix, R is correlations matrix and 1 represents a matrix of same size as correlation matrix where all elements equal 1. This transform gives us distance of 0 for perfectly correlated values, 1 for uncorrelated values and 2 for perfectly anti-correlated ones. Next, the distance matrix is clustered using complete linkage method, designed to find similar clusters. This clustering is used to order and organise the heatmap. For better and more intuitive visualization, I used a gapped heatmap, where distance between the tiles is proportional to distance between values in distance matrix (in classical heatmap tiles are adjoined or separated by a fixed distance).

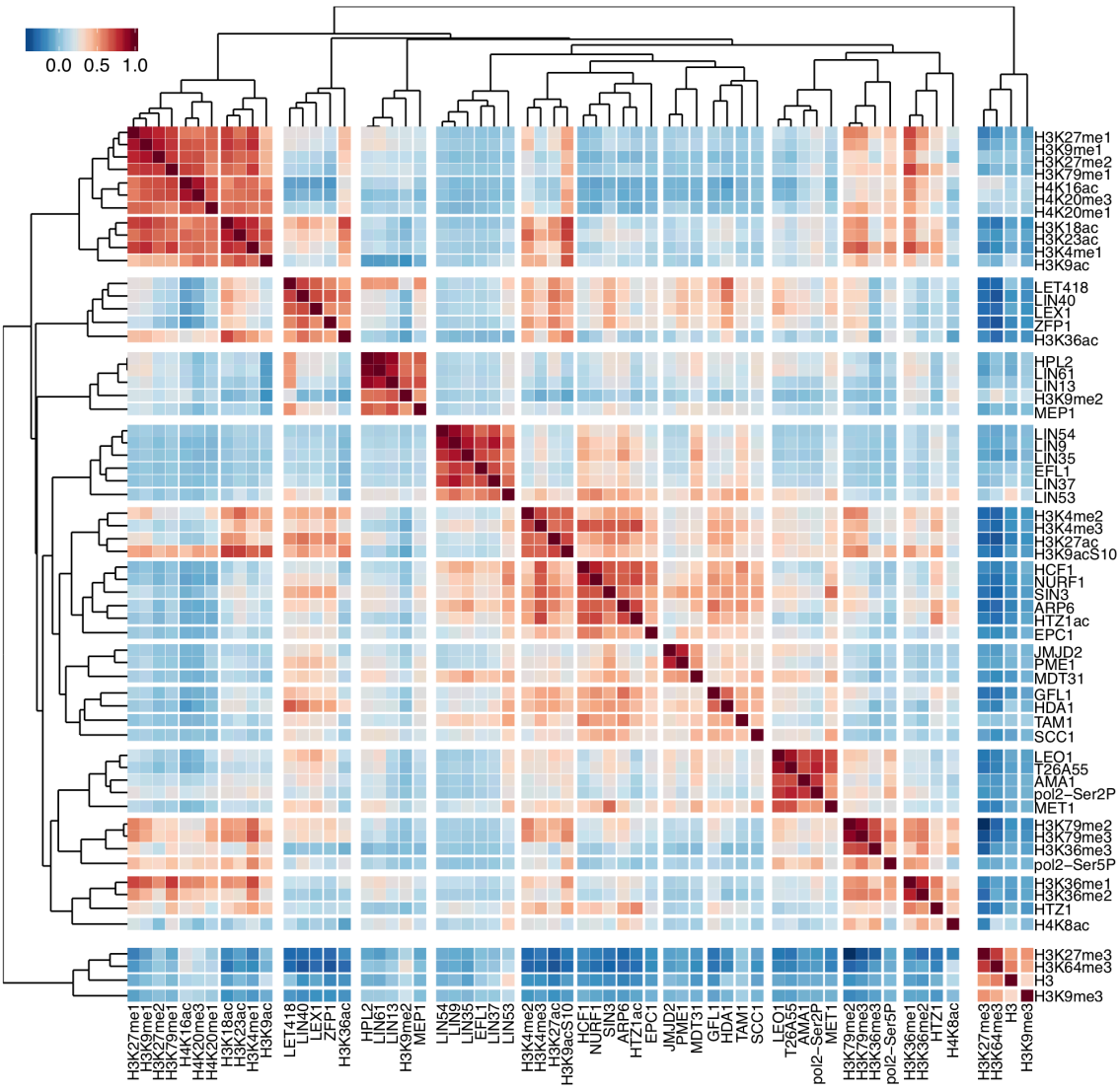


Figure 99 Clustered correlation heatmap at 1KB resolution for 61 factors immunoprecipitated in wild type in L3 larva stage. Tiles colour represents correlation strength, where blue denotes anti-correlated experiments, and red – positive correlation. Dendrograms at top and left edges of the heatmap represent hierarchical clustering based on correlation.

The data acquired at 1kb distance resolution reproduced the clusters found in PCA/k-means analyses (**Figure 99**). Moreover, this figure reveals much more complicated structure within the data, with a number of strong clusters, and numerous further interactions between them. Similarly to PCA, the strongest and most distant cluster is formed by H3K9me3, H3K27me3, H3K64me3 and H3 – this strongly heterochromatin cluster is anti-correlated to all other clusters in the dataset, with rare exceptions, like H3K64me3 being correlated to H3K9me3 or H3 to LIN-53. The largest cluster on the upper left of the heatmap contains many different histone modifications associated with

gene activity such as H3K27me1, H3K9me1, H3K20me, etc. There is a strong HC factor cluster, containing HPL-2, LIN-61, LIN-13, H3K9me2 and MEP-1. Interestingly, LET-418 is still well correlated with these factors, but due to even better correlation it was grouped with LIN-40, LEX-1, ZFP-1 and H3K36ac. Intriguingly, in this group only LET-418 is correlated with HC factors, while other members are anti-correlated. This supports a model presented in Chapter 3, that assumes LET-418 having a yet another role, not connected to HC or genomic repeats suppression. Further, there is a very strong DREAM complex (Latorre *et al.* 2013) containing LIN-54, LIN-9, LIN-35, EFL-1, LIN-37, and LIN-53. Proteins in this cluster are generally correlated to promoter regulatory complexes, and particularly with factors such as HCF-1, ARP-6 and MDT-31 and TAM-1. Next, we can observe a strong promoter cluster containing H3K4me3, H3K4me2, H3K27ac and H3K9acS10, closely connected to a regulatory/histone modifier cluster containing HCF-1, NURF-1, SIN-3, ARP-6, HTZ1ac, EPC-1. Next, we have two small clusters, strongly correlated with previous two. First one contains JMJD-2, PME-1, MDT-31, and second GFL-1, HDA-1, TAM-1 and SCC-1. These four clusters form a large well correlated group, which is constituted of factors present at promoters and enhancers. Further, we have a very tight cluster connected with transcription initiation and elongation – it contains AMA-1 (RNA polymerase 2 globular domain), Pol II Ser2P, MET-1, LEO-1 and JHDM-1 (T26A5.5) (He *et al.* 2008). MET-1, LEO-1 or JHDM-1 have no reported role in initiating transcription process, but they might be important for regulation or stabilization of gene expression. Interestingly, the two histone modifiers present in this cluster have opposite roles - MET-1 is responsible for H3K36me3 deposition and work co-transcriptionally, while JHDM-1 is H3K36 demethylase. Their close connection to transcription might explain co-localisation with Pol II. This illustrates an important feature of machine learning analyses of big genomics datasets – they can reveal interesting interaction in dataset, but good correlation does not imply similar function - actually some factors can have

antagonistic roles. Next, we have a cluster related to gene body associated modifications or factors – H3K36me3, H3K79me2, H3K79me3 and Pol II Ser5P, and a closely related cluster of weaker gene body related marks - H3K36me1/2, HTZ-1 (Latorre *et al.* 2013) and H4K8ac. We can also see some interesting individual relationships – for example, H3K9ac is strongly anti-correlated to HC cluster, especially with H3K9me3.

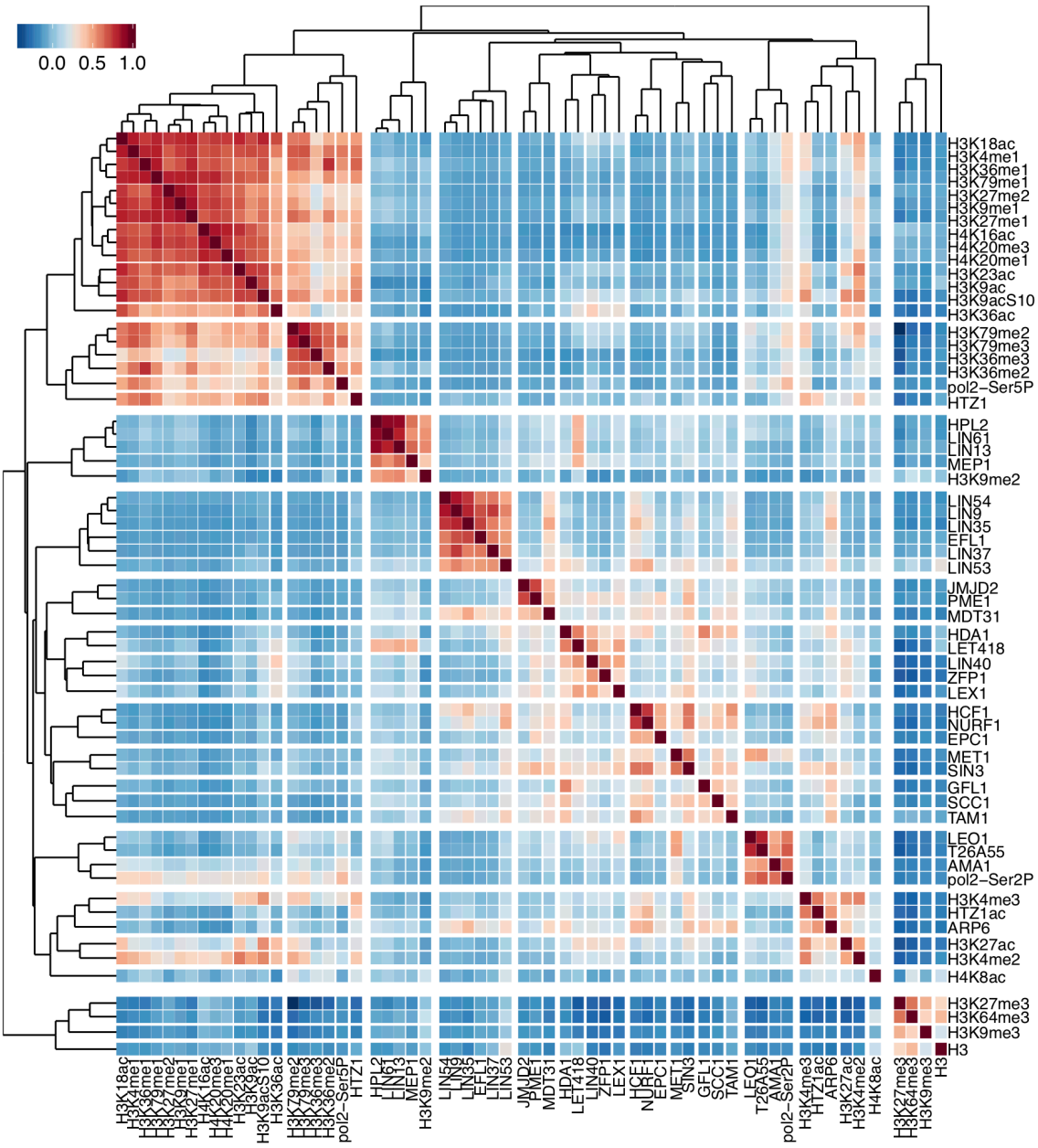


Figure 100 Clustered correlation heatmap at 100bp resolution for 61 factors immunoprecipitated in wild type in L3 larva stage. Tiles colour represents correlation strength, where blue denotes anti-correlated experiments, and red – positive correlation. Dendrograms at top and left edges of the heatmap represent hierarchical clustering based on correlation.

The heatmap based on 100bp resolution (**Figure 100**) shows similar cluster composition, with slightly different configuration of factors constituting each cluster. Local clusters seem more tightly correlated, but most of the global cluster to cluster interactions are not visible any more, and most factors are anti-correlated outside of their local clusters. However, this allows us to clearly see some individual interactions between distant clusters – for example H3K9acS10 is strongly correlated with H3K9me2, and H3K27ac. Similarly, I observe ARP-6 is correlated with SIN-3, NURF-1, HCF-1, HDA-1 and LIN-35.

In conclusion, studying the structure of correlation between factors in ChIP-seq profiles datasets can lead to finding interesting interactions outside of known complexes. However, these interactions do not implicate causality or functional synergy and require experiments to determine if correlation in the profiles indicates some functional or mechanistic relationship.

4.5.6 Graphical model estimation

To further analyse the correlation and provide more intuitive visualization of profile correlation-based relations I decided to investigate the utility of graphical models. This method takes the correlation matrix and represents it a connected network graph, where nodes indicate factors of interest and edges the interactions based on correlation values. Green edges indicate positive correlations and red edges - negative ones. The width of the edges and the colour saturation corresponds to the absolute value of correlation and scale relative to the strongest weight in the graph. The graphs are organised as a "spring" layout, which uses the Fruchterman-Reingold algorithm (Fruchterman & Reingold 1991) to obtain a force-directed layout. In this solution each node (connected and unconnected) repulses each other, and connected nodes attract each other. After a number of iterations (500) a final layout is reached – the distance between the nodes

correspond well to correlation between the nodes – correlated nodes are close to each other, while anti-correlated (negative correlation) are moved to distant parts of the graph.

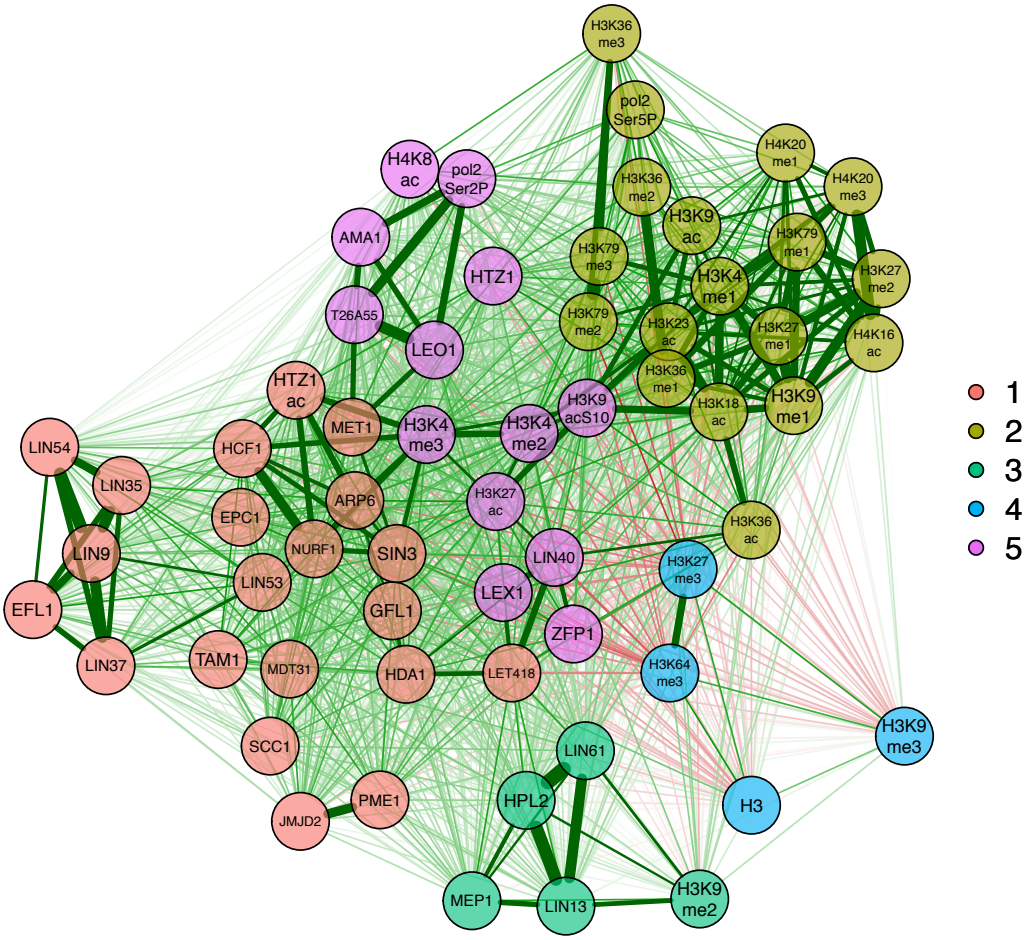


Figure 101 Correlation based Gaussian Graphical Model at 1KB resolution for 61 factors immunoprecipitated in wild type in L3 larva stage visualised as network. Nodes represents factors, while edges represent correlations. Edge width is proportional to absolute correlation value and edge colour represents the direction of interaction, where red dentate anti-correlated experiments (negative relation) and green positive correlation (positive relation). Clusters, derived from PCA analyses, are represented as colour ellipses, and data assignment is colour coded.

On first examination graphical model (**Figure 101**) shows similar structure, as was shown on the cluster heatmaps. To better relate the output of this method to hierarchically clustered heatmaps and k-means clustered PCA, I coloured the nodes

following 5 cluster colour scheme from 1kb bins, 5 cluster PCA analysis: (1) red cluster represents chromatin regulators, (2) yellow cluster represents gene body/transcription elongation elements, (3) green cluster is heterochromatin/repeat elements associated factors (HC), (4) blue cluster is heterochromatin domain/inactive marks, and (5) violet cluster represents promoter/enhancers/transcription initiation associated factors. The cluster annotation is fully data driven and is shown here to compare graphical models to clustered PCA and give some intuition what the layout of the figures is.

In general, this five-cluster structure is also visible in the graphical model. However, we can learn much more about specific interactions, for example we see distinct, well connected DREAM factor cluster, that with PCA was just a part of bigger regulatory cluster. Also, transcription initiation and elongation cluster with AMA-1, LEO-1 and Pol II Ser2P forms distinctive structure from other factors in enhancer/promoter cluster. Interestingly, we see some examples of interfacing factors – for example HTZ-1 is between initiation and gene body cluster. Indeed, HTZ-1 marks strongly TSS of active genes (in similar fashion to H3K4me3), but also extends to gene bodies (Latorre *et al.* 2013). H3K9me3 is far from any other factors, but H3K27me3 is closer to gene body cluster – and indeed it marks bodies of some regulated genes (Young *et al.* 2011). Also a connection, missed by PCA between LIN-40 and LET-418 - both members of LET-418/Mi-2/NuRD complex (Ahringer 2000; Passannante *et al.* 2010b), becomes well visible in the graphical model. This example shows the advantage of graphical models over clustered PCA.

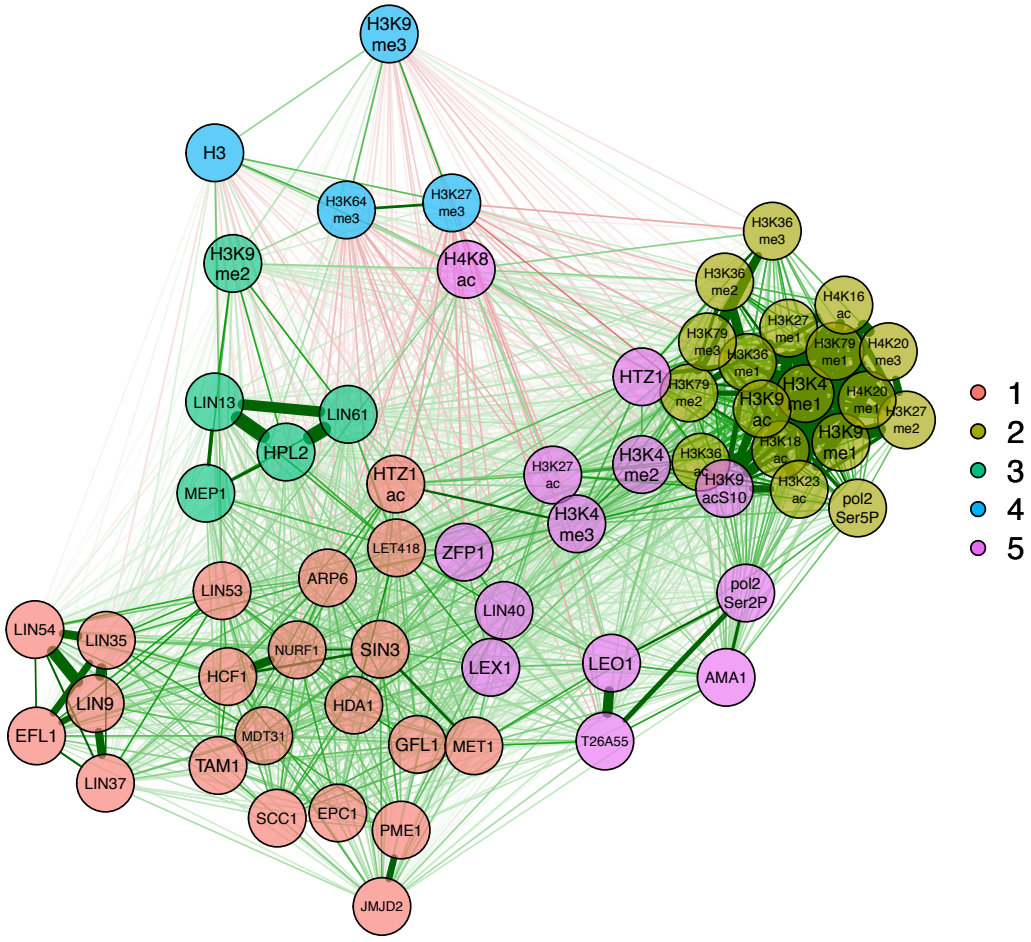


Figure 102 Correlation based Gaussian Graphical Model at 100bp resolution for 61 factors immunoprecipitated in wild type in L3 larva stage visualised as network. Nodes represent factors, while edges represent correlations. Edge width is proportional to absolute correlation value and edge colour represents the direction of interaction, where red denotes anti-correlated experiments (negative relation) and green – positive correlation (positive relation).

When looking at the higher resolution 100bp binning based graphical model, I observe similar trends as for PCA and heatmap correlation analyses – correlation within the primary clusters become stronger, and interactions between groups weaker – for the graphical model this has a consequence of similar groups being clumped together. However, we can observe some interesting relationships within the networks – at lower resolution H3K9me2 moved towards the repressed cluster, interfacing between it and HC factors. We also see a newly formed cluster of SCC-1, EPC-1, PME-1 and JMJD-2.

In conclusion, graphical gaussian models are a useful way to visualize genomic associations, that provide more intuitive way of analysing correlations than heatmaps. However, the structure of the plot is heavily driven by strongest interaction in the network which might be an advantage for assessing the core structure but might mask more subtle interactions.

4.5.7 Graphical models using partial correlation estimation

The correlation based graphical model is strongly driven by strong correlations between well interconnected clusters. This is useful to understand the global structure of the data but can mask more subtle interactions between the clusters. In order to investigate these interactions, I used partial correlation estimation – the method which normalises the strength of interactions to the global structure of data.

The partial correlation matrix is calculated using a sparse Gaussian graphical model with the graphical lasso (Friedman *et al.* 2014). The tuning parameter is chosen using the Extended Bayesian Information criterium (EBIC) (Foygel & Drton 2010; Friedman *et al.* 2014; Higham 2002). EBIC lasso algorithm requires a positive definite matrix as input. Correlation matrices obtained from complete datasets have this property, however the missing values in ChIP-seq data can cause correlation matrices to become not positive definite. In such case the nearest positive definite matrix is calculated using Higham algorithm implemented in “Matrix” package in R. The partial correlation matrixes are then visualised as described in “Graphical models estimation” section. The graphical model based on partial correlations allows us to better understand a global interaction structure of a given gene set, as clusters in the network are not driven by small groups of strong interactions as in graphical models derived from direct

correlation. In other words, correlation analyses reveal associations within the data, while partial correlations give us a concentration or structure of the data.

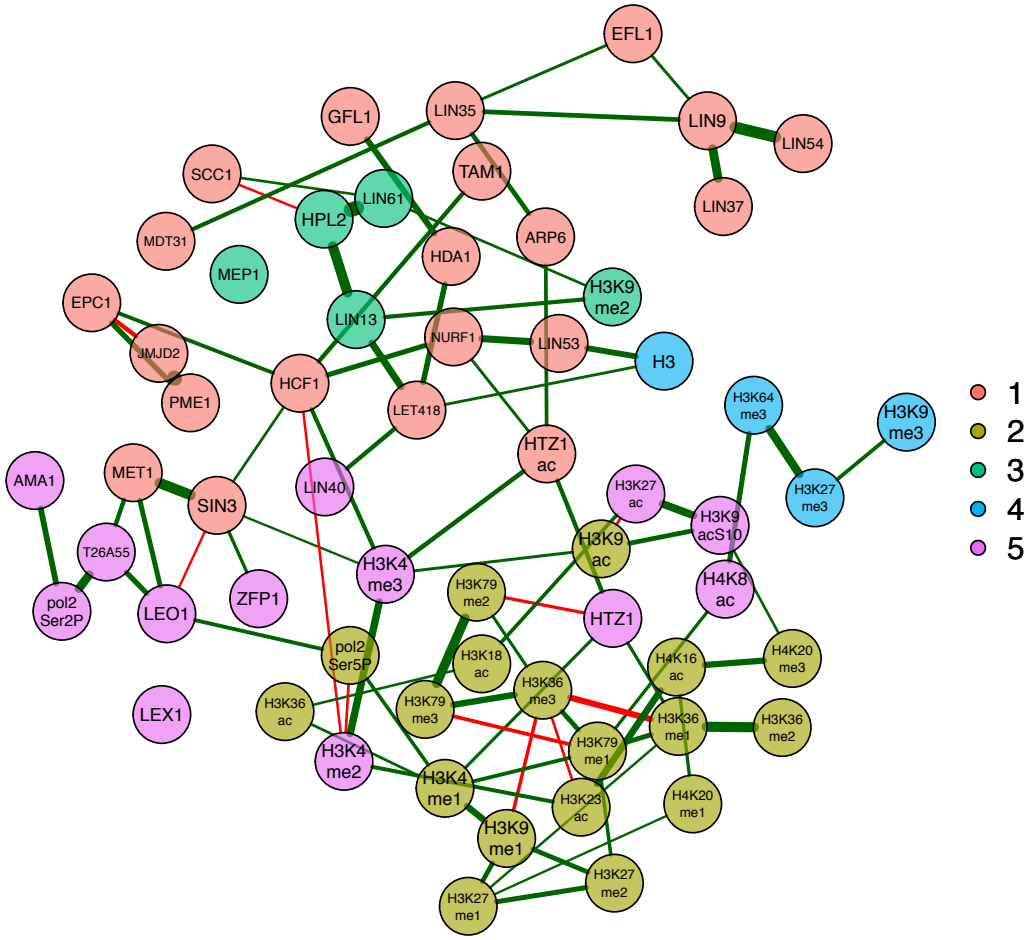


Figure 103 Partial correlation based Gaussian Graphical Model at 1KB resolution for 61 factors immunoprecipitated in wild type in L3 larva stage visualised as network. Nodes represent factors, while edges represent correlations. Edge width is proportional to absolute correlation value and edge colour represents the direction of interaction, where red denotes anti-correlated experiments (negative relation) and green – positive correlation (positive relation). For clarity, only correlations > 0.2 or < -0.2 are shown.

The partial correlation estimation graph at 1kb resolution (**Figure 103**) shows quite different structure than the graph built directly with correlation coefficients. It organises the structure of data, where on horizontal extremes we see inactive state marks (right), and transcription initiation associated marks (left), and on vertical extremes promoter/enhancer regulatory marks (top) and gene body marks (bottom). Also, I see

many negative interactions being much more visible after normalising out strong positive interactions within the clusters. This includes some interactions between me1 and me3 states of some marks, for example H3K79me1 and H3K27me3, and H3K36me1 and H3K36me3. Interestingly, the algorithm estimated partial anti-correlated regulation between H3K79me2 and HTZ-1 – both mark gene bodies, however, HTZ-1 only marks a small subset of gene bodies and it is most often found at promoters (Latorre *et al.* 2013). If they are indeed mutually exclusive it could indicate some possible function in regulation. Also, some TFs are also found to be anti-correlated, like JMJD-1 and EPC-1, as well as HPL-2 and SCC-1. Testing if these predictions have any biological meaning requires further investigation.

In contrast to direct correlation analyses, the partial correlation graph in 100bp resolution resembles the one in 1kb resolution. The graph is organised slightly differently, with better separation of individual groups, but major correlations and anti-correlation driving the graphical model remained similar. This indicates that partial correlation estimation worked well, removing strong correlations coming from small, tight clusters.

In conclusion it seems that partial correlation analyses are useful, as they allow one to uncover network scale structure and interactions between the data, rather than focusing on local, well connected and correlated clusters. However, since the correlation values here are not calculated from the data, but rather estimated, every conclusion driven from this analysis have to be tested and supported by other experiments. Nevertheless, graphical models can be a useful utility in building biological hypothesis based on big datasets.

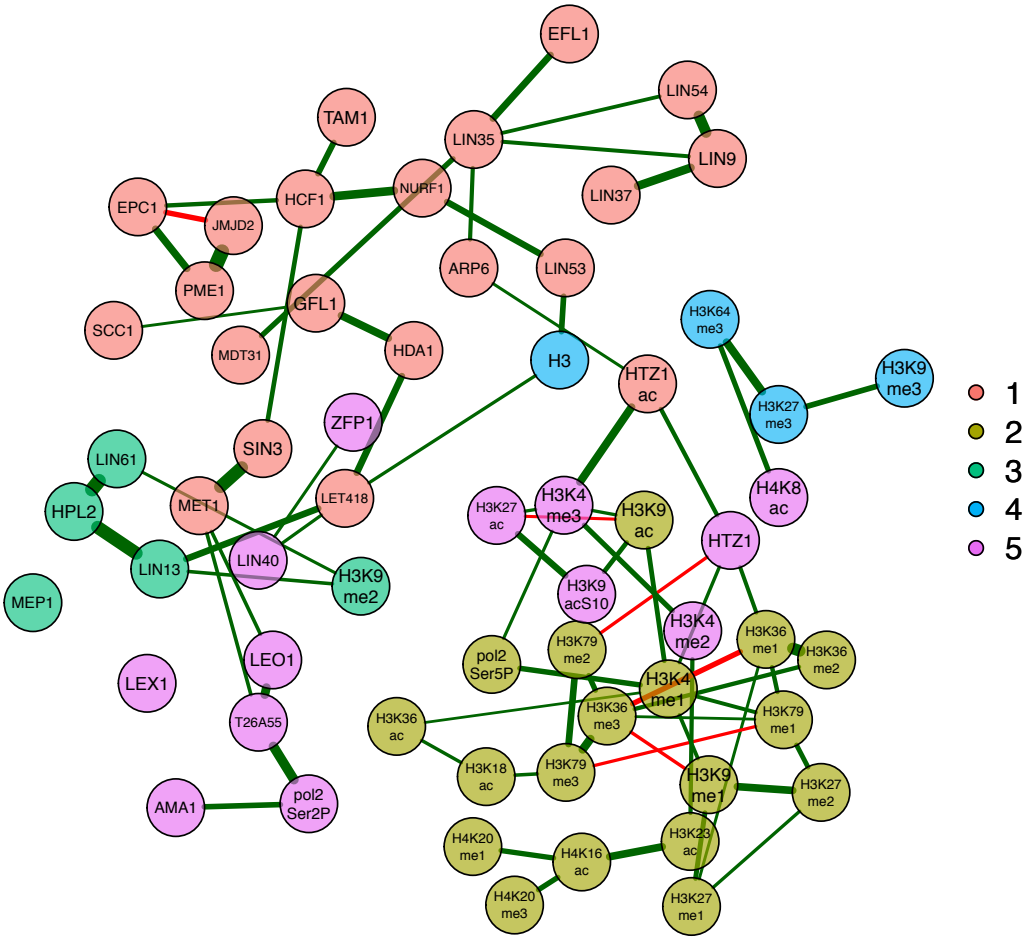


Figure 104 Partial correlation based Gaussian Graphical Model at 100bp resolution for 61 factors immunoprecipitated in wild type in L3 larva stage visualised as network. Nodes represent factors, while edges represent correlations. Edge width is proportional to absolute correlation value and edge colour represents the direction of interaction, where red denotes anti-correlated experiments (negative relation), and green – positive correlation (positive relation). For clarity, only correlations > 0.2 or < -0.2 are shown.

4.6 Matrix factorization method - Nonparametric Sparse Factor Analyses (NSFA)

In order to implement a state-of-the-art pattern recognition method to our analyses I used the algorithm conceived by Zoubin Ghahramani from the Machine Learning Group in the Department of Engineering, University of Cambridge. I applied the nonparametric Bayesian latent feature model – an unsupervised machine learning method (Ghahramani 2004, 2011; Knowles & Ghahramani 2010). In this model an experimental data matrix is modelled as a superposition of a potentially infinite number of hidden features. I implemented Indian Buffet Process (IBP) (Griffiths & Ghahramani 2011) as a prior on superposition matrix. The number of hidden features (factors) is inferred directly from the data. The most significant advantage of this model is that the pattern similarity between profiles is reported simultaneously with genomic positions supporting the patterns. The latent feature models were previously successfully applied to biological studies (Kirk *et al.* 2012; Knowles & Ghahramani 2010); however in those studies the algorithm was run on a small number of samples and positions (about 100 of each) (Knowles & Ghahramani 2010). In contrast, I utilised NSFA to analyse >100 samples with 100 million positions, binned to 1kb (around 100 000 data points). Therefore, I modified the algorithm and implementation to handle our data efficiently.

In general factor analyses model goes as follow:

$$y_n = G * x_n + \varepsilon_n$$

where y_n denotes observed data, for example a single ChIP-seq experiment profile encoded as vector, G is a factor loading matrix and, x_n are latent factors, and ε_n is a noise vector (**Figure 105**).

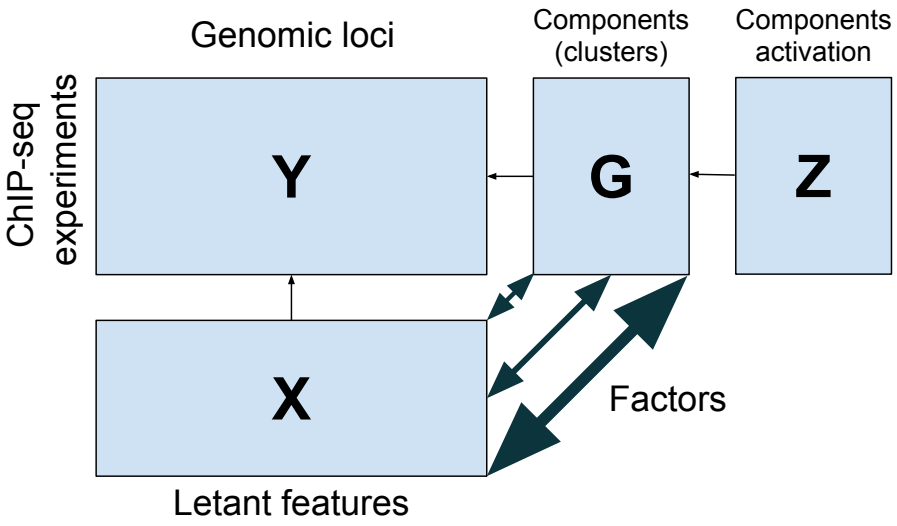


Figure 105 The ideogram shows that matrix Y (observed data, e.g. values from ChIP-seq experiment) is reconstructed by first imposing sparsity (Z) on factor loading matrix G and them multiplying by leant features matrix X.

NSFA extends this model with an activation matrix (**Figure 106**) that is a sparse binary matrix, which decides which loadings are important for the model, hence facilitating sparsity in the model.

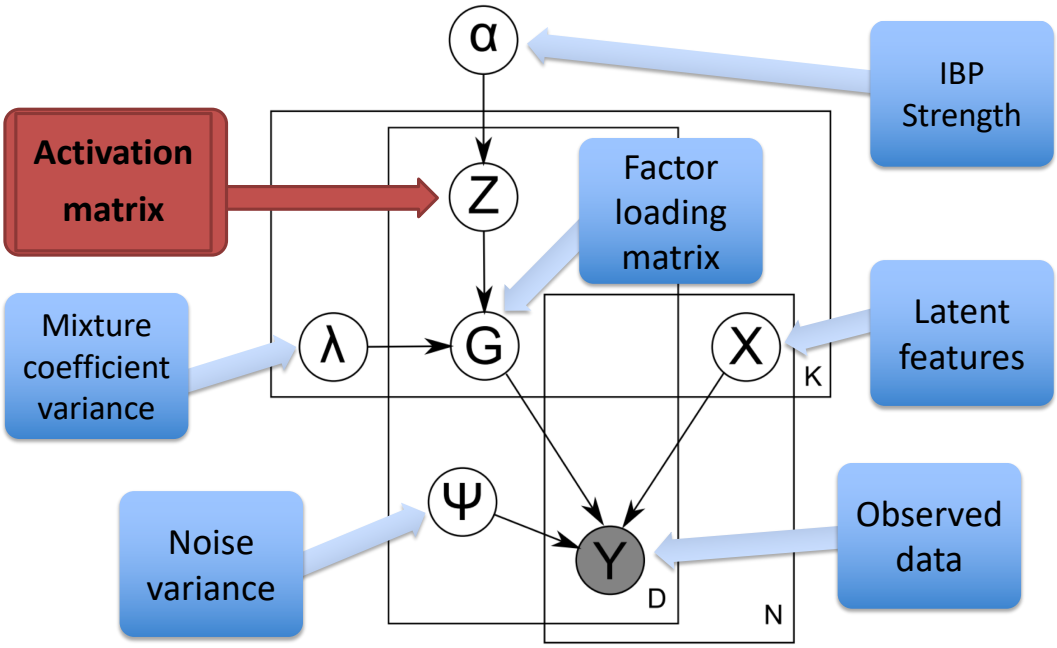


Figure 106 The ideogram of NSFA model. The dimensionality of data matrix is NxD, while number of factors equals K. Adopted from (Knowles & Ghahramani 2010)

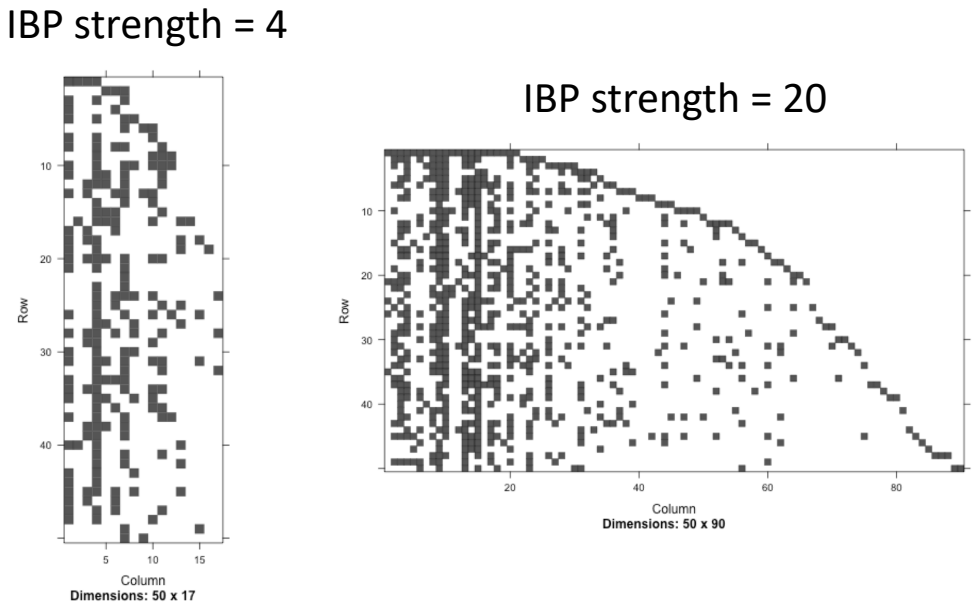


Figure 107 Examples of activation matrix prior established with Indian buffet process (IBP) using two different strength parameters – 4 and 20. In case of strength parameter equals 4 the prior model assumes 17 factors, and with strength equals 20 – 90 factors.

This matrix is in turn controlled by IBP strength hyperparameter, which controls how many factors are in use (**Figure 107**). In addition, there are 2 parameters, whose values are established using sampling methods – mixture coefficient variance, which controls factor loading matrix and noise variance, which captures the noise.

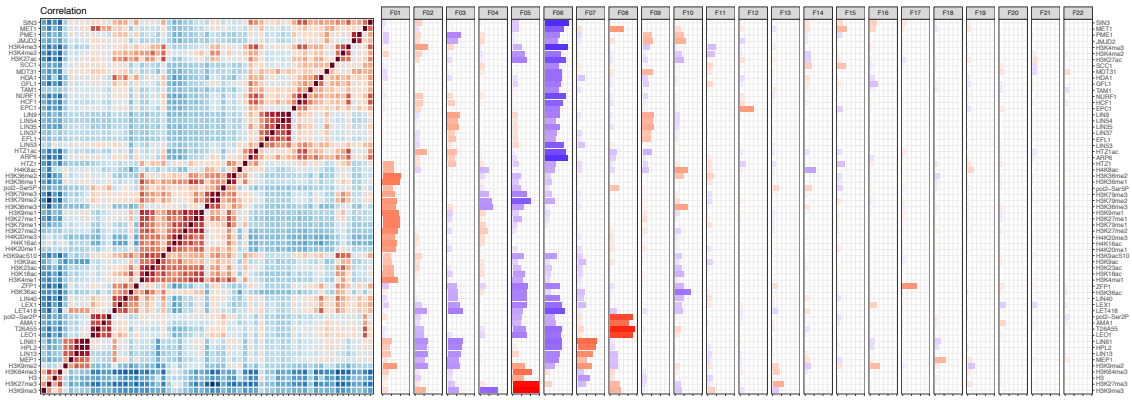


Figure 108 NSFA using all chromosome data at 1kb binning resolution. The right panel shows loading values (G matrix) for each factor (F01-F21). Final order of data was established by applying hierarchical clustering to loading matrix. On the left site the correlation matrix following the order obtained from clustering G matrix. Please note that the correlation matrix is shown only for visualization – the ordering, hence cluster formation is based purely on NSFA loadings.

I ran my NSFA implementation on the previously used 61 factors mapped in wild type L3 larva. After 100 iterations the model converged to 22 factors with IBP strength parameter of 5.53 and relatively low sparsity of 0.898, meaning ~90% of loading matrix G is used in the model (**Figure 109**). The model produced an interesting data division, in global scale very different than structures obtained by correlation analyses (**Figure 108**). However, on the local scale we see the same strong clusters, as were found by correlation analyses – strong grouping of known complexes (e.g. DREAM complex), HC factors, or inactive chromatin factors. Also, despite the fact that model converged at 22 factors, first 8 factors are driving the most structure in the data. Further factors have strong loading only for single, or small groups of experiments, having a role in fine-tuning the structure, rather than driving a global structure.

.

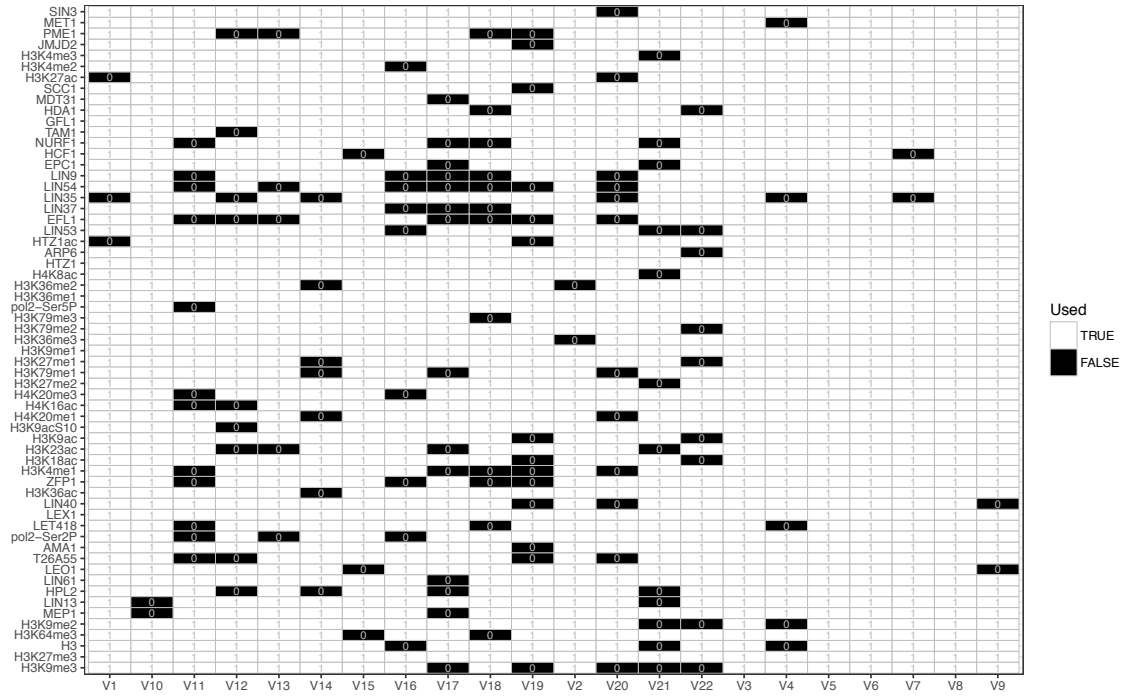


Figure 109 Sparsity matrix derived from NFSA using all chromosome data at 1kb binning resolution model.

The great advantage of factor analyses over methods like PCA is that factor loadings are driven by best fit to the structure of the data, rather than being designed to capture most variance in dataset. This allows us to interpret factor meaning in an intuitive way and

use the combination of factors to extract information of interest. For example, F05 and F06 contribute most to the global structure of the model, F01 divides the data into 3 major classes – regulatory, gene body and other/specialised factors, and F08 is strongly driven by transcription initiation factors (**Figure 108**). We can use this property to visualise the data in selected context. For example, we can plot F05 vs. F06 as scatterplot (**Figure 110**). This analysis gives us similar structure to the one obtained using PCA analyses. However, to focus on expression initiation group in context of other factors we can plot F01 vs. F08 (**Figure 111**). This gives quite different clustering, with initiation cluster clearly separated on F08 and other experiments separated along F01.

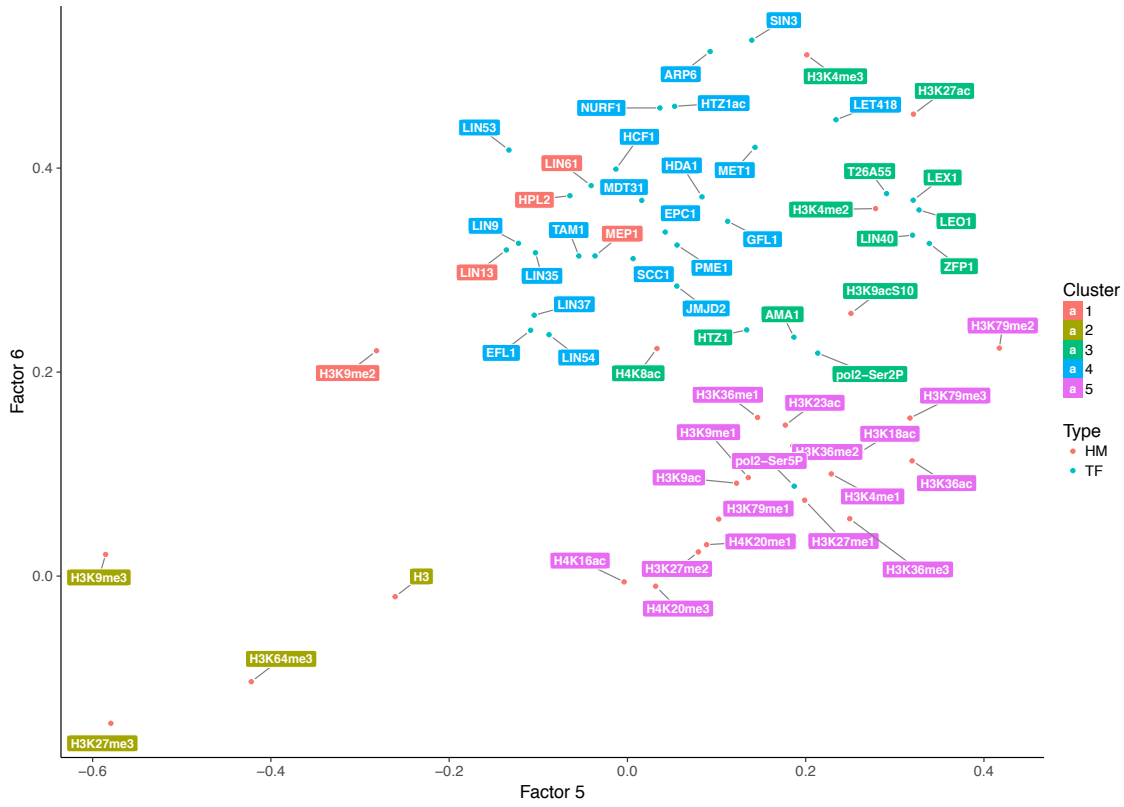


Figure 110 Factor F05 versus factor F06 loading plotted as scatterplot.

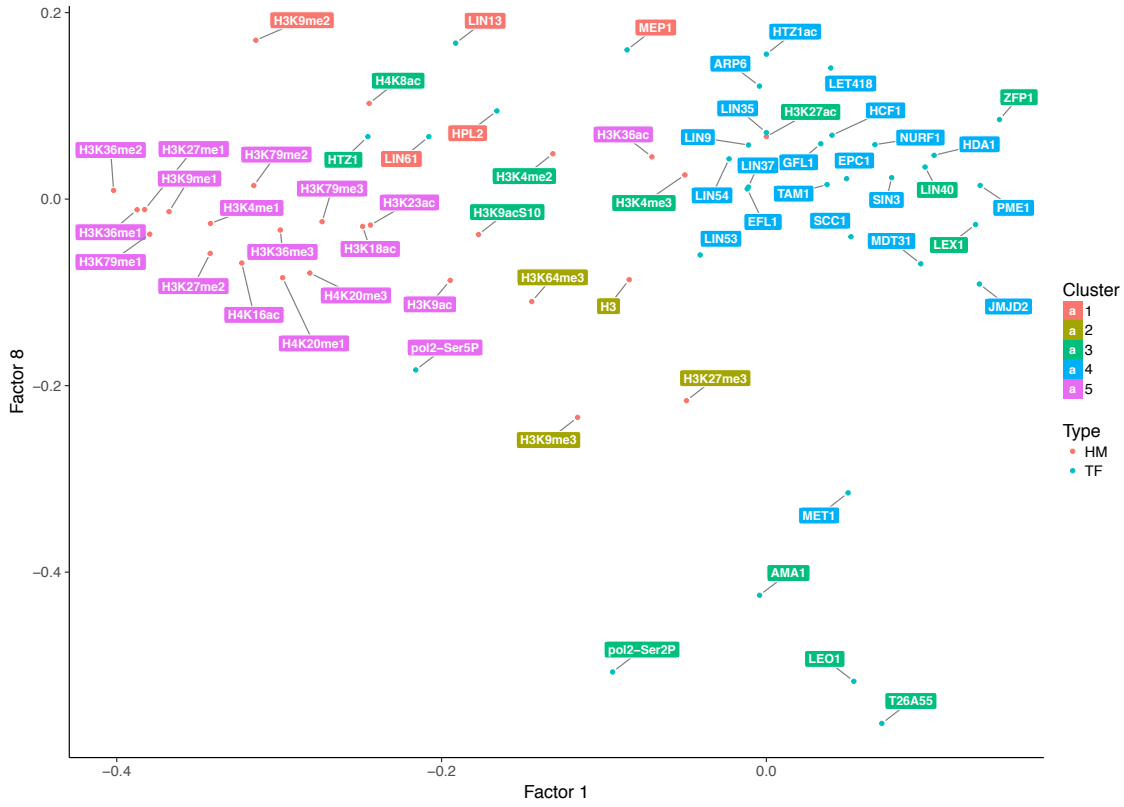


Figure 111 Factor F01 versus factor F08 loading plotted as scatterplot.

Further, I wanted to investigate what is the generalizing power of NSFA. I started with investigating how much variance is captured in each factor, and how much variance in general is captured. As I observed on the factor plot before, most of data variance is captured by factors F05 and F06, and only 10 first factors have a significant contribution to variance captured (**Figure 112**A). Intriguingly, the cumulative value of variance explained by all factors sums up to 32.6%. This mean, that more than two thirds of the variance in our dataset is explained by noise matrix and not incorporated to factors (**Figure 112**B).

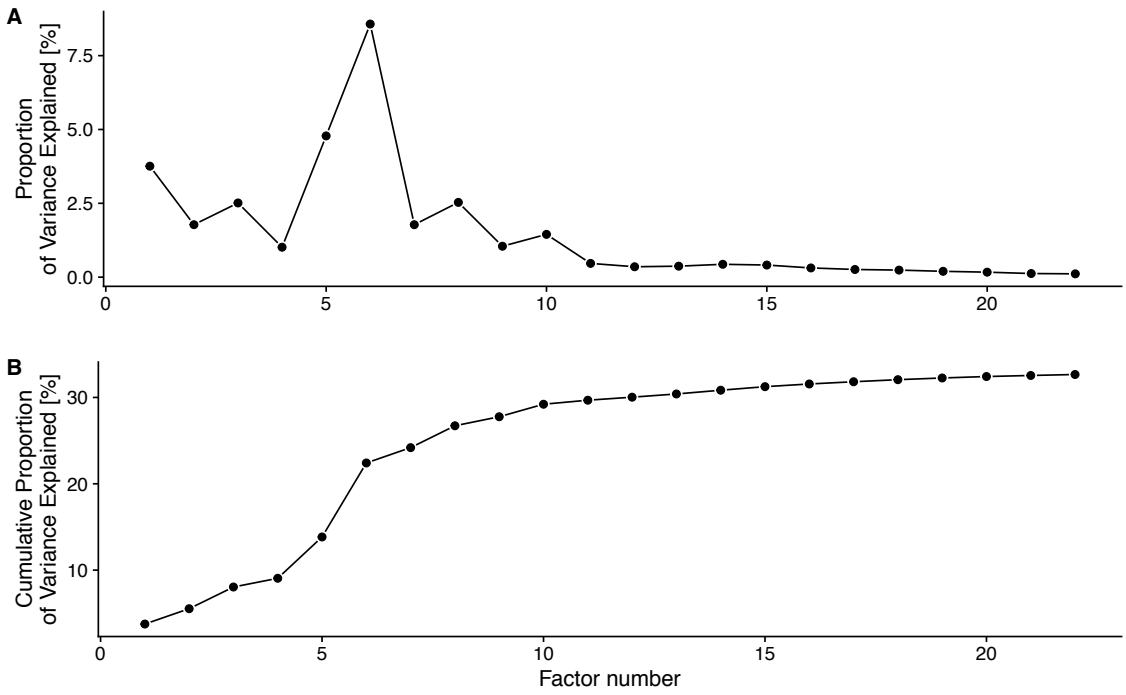


Figure 112 Diagnostic plots for NSFA acquired with 1kb binning resolution. (A) Proportion of the explained variance plot and (B) Scree plot showing cumulative proportion of variance captured by given number of factors.

The obvious question arose – does capturing only ~30% of variance in dataset is enough to properly model interactions within the data? To answer this question, I have produced a graphical model based on NSFA loadings, rather than actual data (**Figure 113**). Astonishingly, the graphical model not only perfectly captured interactions within the data, but global structure is tidier and interaction between known complexes stronger than using any other method tested before. This result is even more impressive, considering, that original data used to produce the graphical model at 1kb resolution had 6,116,409 data elements and sparse loading consisted of only 1208 elements, achieving 5063 times data reduction without detectable loss of informative value, and possibly even a gain, since the noise was efficiently filtered out. In conclusion, NSFA can extract latent structure of ChIP-seq data, remove the variance coming from noise, and greatly reduce the dimensionality of the data.

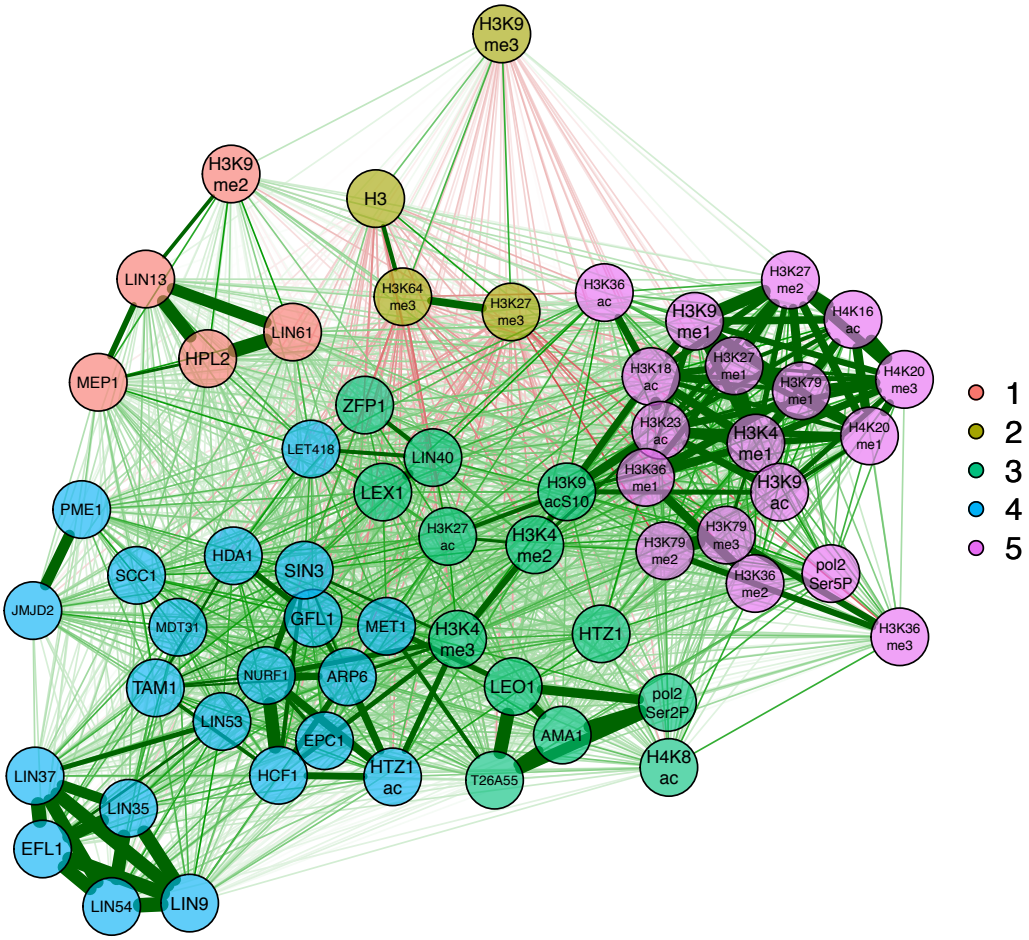


Figure 113 NSFA loadings based Gaussian Graphical Model at 1KB resolution for 61 factors immunoprecipitated in wild type in L3 larva stage visualised as network. Nodes represent factors, while edges represent correlations. Edge width is proportional to absolute correlation value and edge colour represents the direction of interaction, where red denotes anti-correlated experiments (negative relation), and green – positive correlation (positive relation). Clusters, derived from PCA analyses, are represented as colour ellipses, and data assignment is colour coded.

By its design principles, NSFA should be good for generalizing big datasets. To test that ability, I restricted NSFA model to train only on the first chromosome. Then I clustered loading matrix G using hierarchical clustering, and visualised factors along the correlation matrix driven from the whole dataset (all chromosomes). The restricted model was able to properly estimate the structure of the data with this restricted dataset (**Figure 114**). Indeed, it produced the same local cluster and similar global structure to

model run on the whole dataset. This illustrates the ability of NSFA to derive good data generalization in ChIP-seq experiments. This feature can be utilised to train the model on limited, but higher resolution datasets, enabling to run high resolution analyses on large genomes. This can be particularly useful to run NSFA in limited computational resources or shorten the time and memory usage when running on datasets composed of thousands of individual experiments.

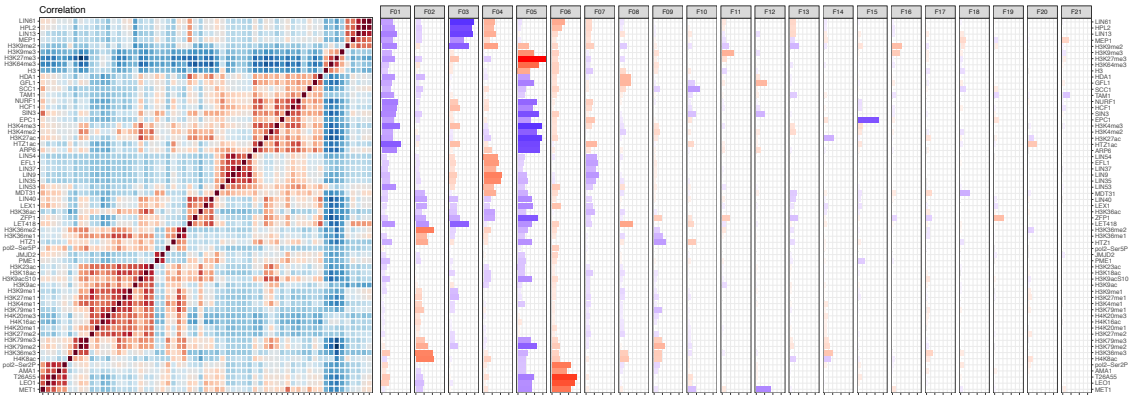


Figure 114 NFSA using ChrI chromosome data at 1kb binning resolution. The right panel show loading values (G matrix) for each factor (F01-F21). Final order of data was established by applying hierarchical clustering to loading matrix. On the left site the correlation matrix following the order obtained from clustering G matrix. Please note, that correlation matrix is shown only for visualization – the ordering, hence cluster formation is based purely on NSFA loadings. Further, the correlation matrix is based on whole genome data, while NSFA is based only on chromosome ChrI – this shows a great potential for generalization in NSFA model.

5 DISCUSSION

In this thesis I present computational exploration of interactions between chromatin features, underlying DNA sequence and gene regulation. I used bioinformatics methods to study heterochromatin factors and active regulatory regions and I performed global analyses of large-scale genomic datasets.

5.1 *C. elegans* as a model organism

In my studies I utilised the advantages of *C. elegans* as a model organism – it has a small (30 times smaller than human), but extremely well annotated genome. This allowed me to perform the computational analyses faster, with less computational resources required. Also, thanks to the collaborative efforts of *C. elegans* community, and consortium projects like modENCODE, there is plethora of publicly available genomic resources. I utilised publicly available datasets in my studies - they enabled piRNA and 22G RNA study in the heterochromatin chapter and were the foundation of the open chromatin, HOT region study. *C. elegans* also has a great advantage for our laboratory use – this model organism is easy to handle, has a rapid generation time and very low-price tag on generating high fidelity sequencing profiles (due to small genome size), which allowed my colleagues to generate data quickly, allowing in turn myself to gather this data and create a vast, uniformly processed collection of high quality ChIP-seq and RNA-seq experiments stored in JADB system. This was a cornerstone for my further investigations – both hypothesis-driven focused studies, and data-driven large

dataset studies, that utilize machine learning algorithms. *C. elegans* shares a very similar promoter architecture with *H. sapiens* – in humans generally similar complexes regulate equivalent histone mark deposition, the only major difference being lack of DNA methylation in *C. elegans*, hence absence of classical unmethylated CpG islands. However, in my studies I have shown that *C. elegans* actually exhibits sharp enrichment of CpGs at promoters at promoters, that in both organisms are recognised by homologues CFP-1/CXXC1 protein, which in both species is required for H3K4me3 deposition. This illustrates a great potential of comparative studies between *C. elegans* and mammals, where findings from each model complement, and one can combine strengths of both models to elucidate on complex biological processes. Further, *C. elegans* is a great model to study the transcriptional control of repetitive elements by factors associated with H3K9 methylation. In higher animals abolishing the writers of this mark causes chromosome mis-segregation. However, in *C. elegans*, possibly due to holocentric chromosomes (Melters *et al.* 2012), and repeats being relatively uniformly distributed on chromosome (with exception of higher repeat density on chromosome arms) no such phenotype is observed. Also, the pattern of dispersed repeats in *C. elegans* means they usually border unique DNA and that reads that cover repetitive – unique DNA junctions can be mapped, allowing us to profile individual repeats in specific loci using RNA-seq and ChIP-seq, rather than being forced to rely on family level bulk analyses. This is especially important, because, as I showed in the genomic repeats profiling study, repeats in the same family are often very diverse. They are derived from a common ancestral repeat, but rapidly diverge in evolution.

5.2 Repeats, heterochromatin factors and germline function

The study began with investigation of the repressed chromatin state. I first showed that a group of heterochromatin factors HPL-2/HP1, LIN-13, LIN-61, LET-418/Mi-2, and H3K9me2 histone methyltransferase MET-2/SETDB1 colocalises. I also showed that

loci bound by heterochromatin factors correlate well with H3K9me2, but not H3K9me3. I found that H3K9me2, but not H3K9me3 is enriched at telomeres. H3K9me3 and H3K9me2 have different genomic profiles - H3K9me3 marks a broad domain, while H3K9me2 shows sharp, transcription factor-like peaks, and they usually do not directly overlap. This might suggest that H3K9me3 in *C. elegans* is a general, broad repressive mark, similar to H3K27me3 (they correlate well, as I show in data driven analyses), while H3K9me2 might regulate specific loci, like small repeats or regulatory regions within a larger region. This trend is well illustrated by a chromosome II Helitron1 repeat cluster (**Figure 10**), where H3K9me2 and all five heterochromatin clusters mark the putative Helitron promoter, while H3K9me3 decorates the Helitron gene body (ORF region). Further, I demonstrated Heterochromatin factor binding loci and H3K9me3 marked loci overlap with repetitive elements. Next, we used the advantages of *C. elegans* model and profiled individual repeat expression, I found that a small number of individual repetitive elements were de-repressed in heterochromatin mutants with MIRAGE1 being up-regulated in all mutants. Remarkably, MIRAGE1 RNAi experiments performed by my colleagues clearly showed that MIRAGE1 de-repression contributes to sterility phenotypes of heterochromatin mutants (McMurchy *et al.* 2017). However, re-silencing MIRAGE1 produced only partial recovery of fertility. Hence, I speculate that the de-repression of other repeats, alongside general impairment of heterochromatin driven repression, might further contribute to the sterility phenotype. Further I investigated the role of piRNA pathway and found that although there are some similarities in de-repressed repeats in HC factors and piRNA mutants, and partial redundancy between the pathways, as shown by *let-418::nrde-2* experiment, the nuclear RNA pathway is more focused towards supressing retrotransposons, while heterochromatin factors repress more DNA transposons.

5.3 Repeats in aging

Further, I found that repeat elements are de-repressed in the soma during normal *C. elegans* ageing. Moreover, this de-repression is consistent – once a repeat was de-repressed in an earlier time point it remained de-repressed. My working hypothesis is that progressive loss of H3K9 methylation might be a cause for repeats misexpression. The discovery that repeats are activated in non-random manner prompted me to ask what is the mechanism that regulates its expression. This hypothesis is based on the studies presented in this thesis, that shows the most likely causes of repeat de-regulation are (1) the reduction of H3K9 methylation, (2) the loss of chromatin factors, and (3) loss of nuclear RNAi pathway functions. From these three possibilities reduction of H3K9 methylation seems the most likely to be a progressive process. To test this model, we tried performing ChIP-seq on ageing wild-type animals, however due to technical difficulties in preparing extracts from old worms, we have not yet obtained conclusive results.

5.4 Heliton1 as model for functional study of TACBGTA motif

I also found that the TACBGTA motif is particularly enriched on repeats and heterochromatin factor binding sites. Searching motif databases, I found that in mouse there is a homologous protein to *C. elegans* proteins ODD-1 and ODD-2 that binds this motif. ODD-1 and ODD-2 are not well characterised in *C. elegans*, with only known function being in gut development. They are zinc finger proteins, similar to LIN-13, and an abundance of their putative binding motif is a good indicator they might be involved in repeat expression. In my opinion, Helitron1 is a great model to study the connection between this protein, heterochromatin factors and repeats mis-regulation. Helitron1 has a single, but strong peak of all 5 HC factors, well co-localised with H3K9me2 peak, and a single TACBGTA motif site. Strikingly, this locus overlaps the putative Helitron1

promoter. Moreover, Helitron1 is weakly up-regulated in some heterochromatin mutants.

It should be noted that Helitron is a rolling-circle transposon, and its transposition mechanisms is different from MIRAGE1. MIRAGE1 is a classical cut-and-paste transposon, which generates double strand breaks (DBS). Helitron is a copy-and-paste transposon, and it produces only single strand nicks, leaving the donor site intact. In *C. elegans* the Helitron encodes only two activities – helicase and nickase. Upon transposition, the nickase cuts one strand of DNA, and helicase unwinds DNA until the termination motif (usually forming hairpin-like structure) is met. Then the second nick is created, and the single strand of DNA can be transposed to an acceptor site. Since the missing strand is promptly repaired (likely repair process is parallel to helicase activity), no DNA lesion is created in donor site, and genotoxic stress is minimal in comparison to DBS. However, upon transposition the nickase cut the acceptor site and the Helitron sequence is inserted, creating a single-strand DNA loop. This loop has to be resolved (either by excising overhanging DNA or synthesizing second strand) during DNA replication for a cell to progress throw cell cycle. This might cause replication stress. Taking this together, I have conceived a model, where upon loss of the repressive function of heterochromatin factors Helitron repeats are activated, which causes transposition events and single strand overhangs. This in turn stalls replication fork, causing replication stress that activates apoptosis pathway in germline and ultimately contributes to sterility. This model is visualised on **Figure 115**. This, yet untested model acts in parallel to the MIRAGE1 contribution to sterility phenotype I have shown in charter 2.

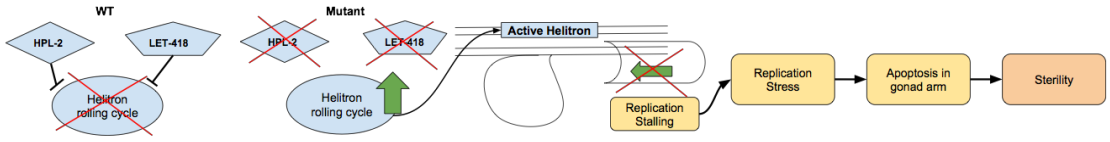


Figure 115 The Helitron induced sterility model: in WT, repeat binding proteins like HPL-2 and LET-418 suppresses the expression of autonomous Helitron repeats. In absence, Helitrons rolling cycle becomes activated, which causes replication stress, which in turn leads to increased apoptosis in gonad arms and sterility phenotype.

Interestingly, Helitron repeats are present in many animals, including mammals. In human there are no active Helitrons, but our genome harbours residues of many past transposition events of Helitrons. This is in contrast to MIRAGE1, which are young, and, as far as we know, nematode-specific elements. This might also explain why MIRAGE1 is the only element de-silenced in all heterochromatin factor mutants – as this repeat is a recent acquisition in *C. elegans* genome, its silencing mechanism may be still evolving.

Moreover, Helitrons are speculated to be an important driver of evolution since they hold potential to duplicate genes and "shuffle" the genome. The helicase termination motif is sometimes lost, or not recognised, which gives Helitrons the potential to harbour downstream sequence and copy it to different genomic loci. If this sequence contains a gene or regulatory element it can trigger the process of the element or gene diverging into two different ones. Interestingly the only branch of mammals with active Helitrons are bats, which is also the fastest evolving species. It was also shown, that by transferring bat's active Helitron to human HeLa cell line normally inactive remnants of Helitrons can be activated (Grabundzija *et al.* 2016).

Finally, I investigated a possibility of splicing defects in HC mutants, since both HC binding sites and repeats often reside within the introns of coding genes. I found no evidence of differential splicing in *hpl-2* mutants, suggesting that HC binding sites and repeat placement in introns do not have a functional role in splicing. However, differential splicing seems to be a quite rare event in *C. elegans*, and hence might be hard to detect (Tan & Fraser 2017).

5.5 CFP-1, HOT sites and CpG dense regions regulate expression of downstream genes

After characterising a suppressing role of heterochromatin factors on repeats I moved to characterise active, open chromatin, regulatory regions. I used a large collection of modENCODE ChIP-seq transcription profiles to define a highly occupied target (HOT) sites in *C. elegans*. Then we showed that the regions are promoters, combining computational analyses and reporter assays. For further analyses I defined COLD regions – loci opposite of HOT regions, where only single TF was bound. I showed that these regions are dense with CpG dinucleotides, which was somehow surprising, since there is no DNA methylation in *C. elegans*. Then I showed that CFP-1/CXXC1 binds CpG dense, nucleosome depleted promoters and, along with SET-2, is required for H3K4me3 deposition at these loci. This function is conserved with its mammalian homolog, though control of CFP-1 binding is different in these two organisms – in humans its binding activity may be controlled by DNA methylation, since it was shown that CXXC1 binds specifically to unmethylated CpG islands. Further I determined that H3K4me3 enrichment correlates with directionality of expression, while depletion of CpG around CFP-1 peak is anti-correlated with elongation directionality. This might potentially contribute to the transition between initiation and elongation of polymerase II and selection of productive elongation direction. Interestingly, bi-directional promoters show a symmetric pattern of both H3K4me3 enrichment and CpG profile. Also, the tri-nucleotides CGC/GCG show even stronger enrichment at promoters, and their imbalance might further contribute to the directionality. Further investigation is required to determine if this mechanism indeed contributes to elongation directionality selection.

Using expression profiling I determined that the majority of CFP-1 binding targets are not significantly mis-regulated in *cfp-1* mutants but are weakly upregulated in bulk

analyses. This stands in contrast to stereotypical view of H3K4me3 and its facilitators – CFP-1 and SET-2 as “transcription activators”, where one would expect a significant drop in expression in mutant backgrounds. I have also shown that CFP-1 driven H3K4 methylation might instead stabilise its target expression, since genes with promoters marked by CFP-1 tend to be more stable in embryo development, as measured by lower coefficient of variance than other classes of tested genes.

5.6 Interactions between CFP-1 and Sin3S/HDAC histone deacetylase complex

Further I analysed interactions between CFP-1 and the histone deacetylase complex Sin3S/HDAC, that includes HDA-1, SIN-3, ATHP-1 and MRG-1. My colleagues generated HDA-1 and SIN-3 ChIP-seq profiles, and I showed that these factors often co-localise with CFP-1. Next, I showed that CFP-1 functionally interacts with HDA-1and SIN-3 using experiments in mutant strains. In the *cfp-1* background I observed both loss and gain of SIN-3 binding, depending on chromatin context. In sites that showed a significant loss of H3K4 trimethylation in *cfp-1* mutant there is also a loss of SIN-3, while in other CFP-1 and SIN-3 binding sites there is a gain of SIN-3. HDA-1 is also lost in sites that lose H3K4me3, but it is unaffected at other loci.

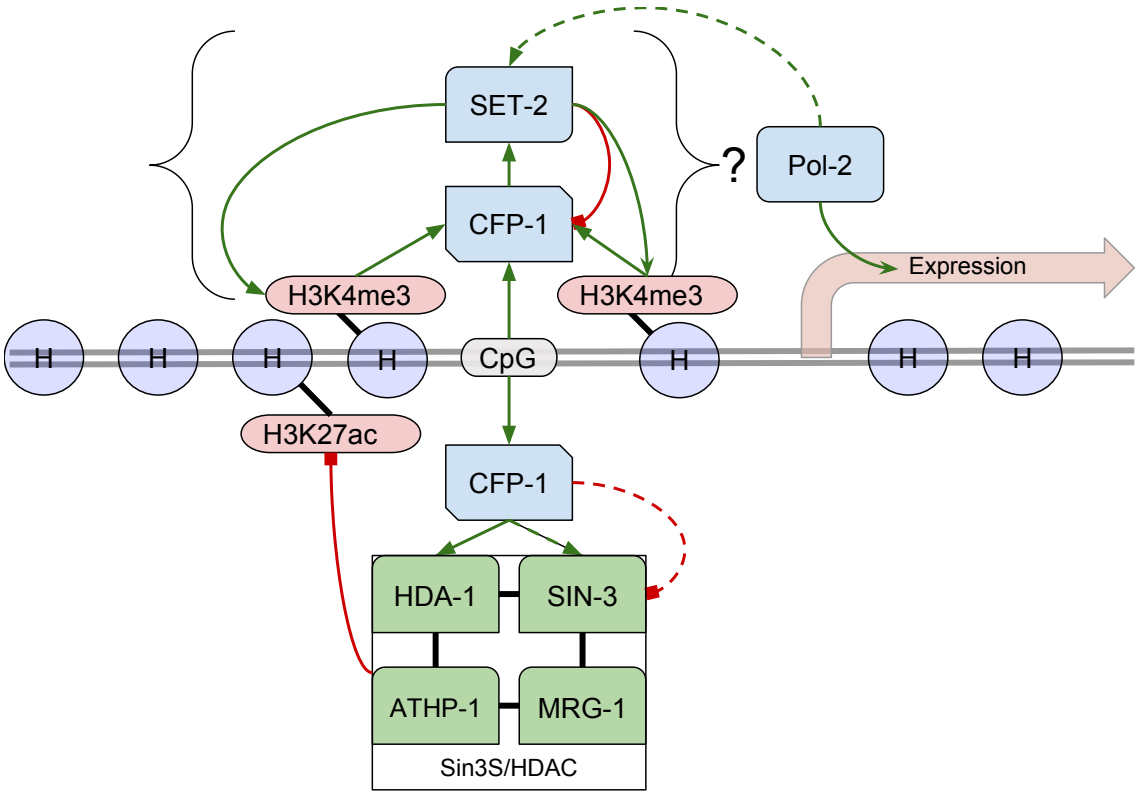


Figure 116 Model of interactions between Sin3S/HDAC, SET-2/COMAPSS and CFP-1 in context of histone modifications and transcriptional activity.

Based on these observations I propose a model, where CFP-1 binds to CpG rich regions in *C. elegans* genome, facilitating further binding of the SET-2/COMPASS and

Sin3S/HDAC complexes. In mammals CFP-1 binding is further reinforced by deposition of H3K4me3 but suppressed by interaction with SET-2. In *C. elegans* CFP-1 lacks the H3K4me3 binding PHD finger, so this connection might not be present. However, it could bind H3K4me3 by an unknown mechanism though there is no evidence one way or another. The negative interaction between SET-2 and CFP-1 comes from study performed on homologues proteins in mouse model, where in the absence of functional SET-2/Set1 there is a vast increase in CFP-1 binding (Brown *et al.* 2017a). Also, transcriptional activity may contribute to a more open chromatin state, further increasing the potential of CFP-1 binding. At the same time, CFP-1 promotes binding of the Sin3S/HDAC complex, which deacetylates histone H3K27. However, at the loci without SET-2/COMPASS, CFP-1 rather prevents binding of SIN-3, protecting H3K27ac marking. Understanding the mechanism of this dual interaction between SET-2 and SIN-3 will require further studies.

In the literature, there are many observations that postulate functional connections between the heterochromatin and the Sin3S/HDAC and SET-2/COMPASS complexes. For example, a synthetic multivulval (SynMuv) phenotype caused by mutations in HPL-2/HP1 and LET-418/Mi2 can be supressed by inactivation of CFP-1, WDR-5.1 or DPY-30 - subunits of SET-2/COMPASS complex, or by inactivation of SIN-3 or MRG-1 - subunits of the Sin3S/HDAC complex (Cui *et al.* 2006; Yücel *et al.* 2014). Further, mutations in DPY-30, WDR-5 MRG-1 and SIN-3 were also shown to suppress the larval lethality in *lin-35* mutant background (Fay & Yochem 2007). These, alongside the data presented in this thesis might suggest CFP-1 being a hub protein that brings together different factors to regulate gene expression.

5.7 Regulatory motifs are important for activation and suppression of transcriptional activity

In my work I have studied two motifs that are particularly connected with regulatory regions in *C. elegans* – an E2F-like motif that is mostly driven by CpG enrichment and TACBGTA – a heterochromatin factor/repeat motif that may attract the binding of heterochromatin factors. They often co-localize in loci marked by heterochromatin factors. It suggests that TAC(B)GTA motif facilitates HC factor binding, which provide active suppression of expression activation. Besides suppressing expression, heterochromatin factor binding and H3K9 methylation marking may have a structural role of preventing homologous recombination in repetitive regions. In the absence of TAC(B)GTA, E2F/CpG motif may be recognized by SET-2/COMPASS and/or E2F/EFL-1 which results in activating an expression programme and putting stabilizing histone mark H3K4me3, deposition of H2A.Z/HTZ-1 and possibly participating in gene regulation.

5.8 Computational methods

During my work I developed a collection of computational tools and method to facilitate processing, storage, retrieval, annotation, data driven and exploratory analyses, visualization, and conceptualization of large datasets in genomics. Two of these tools – SeqPlots and rBEADS are publicly available, and others are accessible to our lab members and collaborators. In future I would like to release more of these tools, as I believe they are beneficial for the scientific community. I have got very positive feedback on SeqPlots.

I also used non-parametric sparse factor analyses (NSFA) to perform a data driven study on a large collection of ChIP-seq profiles. NSFA has a number of advantages over similar computational models, making it particularly well suited for genomics data

(**Table 22**). Further I compared it to other unsupervised machine learning algorithms. In my opinion I demonstrated the utility of this method, as it uncovered interactions and learned structure in the genomic dataset. The next step will be applying it to more dynamic states, like wild type - mutant experiment pairs or different developmental stages to uncover more interesting interactions. In my opinion, unsupervised machine learning algorithms will help us to better understand large and complex biological datasets.

Feature	CA	HMM	PCA	NSFA
Dimensionality reduction	+	+	+	+
Interactions between proteins	+	-	+	+
Interactions between genomic loci	-	+	+	+
“Learns” data structure	-	+	-	+
Noise model	-	-	-	+
Sparsity model	-	-	-	+
Nonparametric	+	-	-/+	+
Simple visualization	+	-	+	+

Table 22 Comparison between NSFA and other models machine learning methods: CA - Correlation Analyses, HMM - Hidden Markov Models & PCA - Principal Component Analysis. “+” means that given feature is present, “-” means it absent and “-/+” means it is partially present.

6 MATERIALS AND METHODS

6.1 Methods used in heterochromatin and repetitive elements study

6.1.1 Alignment to reference genome for ChIP-Seq and RNA-Seq data

Chip-seq and RNA-seq libraries were sequenced using Illumina HiSeq. Reads were aligned to the WS220/ce10 assembly of the C. elegans genome using BWA v. 0.7.7 (Li and Durbin, 2010) with default settings (BWA-backtrack algorithm). The SAMtools v. 0.1.19 'view' utility was used to convert the alignments to BAM format. To be able to investigate binding and expression at repetitive elements, we used all aligned reads (mapq0) to generate pileup and normalised tracks. Normalized ChIP-seq coverage tracks were generated using the BEADS algorithm (Cheung et al 2011), without mappability correction step.

6.1.2 Summed ChIP-seq *Input* and in-house blacklist

We generated summed input BAM files by combining good quality ChIP-seq input experiments from different extracts (8 experiments for formaldehyde and 5 experiments for EGS extracts). The same summed inputs were used for BEADS normalisation and peak calls. We observed that despite using input files for Model-based Analysis for ChIP-Seq peak caller (MACS2) (Zhang *et al.* 2008) and filtering against

modENCODE, some regions of high signal in input were still called as peaks. To overcome this problem, we created an in-house blacklist by running MACS2 with default settings and no input mode. The blacklist regions were refined by discarding region with MACS2 score below 100 and clustering peaks within 500bp. This procedure created 90 new regions in addition to 122 already covered by modENCODE blacklist.

6.1.3 Peak calls

Broad and sharp ChIP-seq peaks were generated as follows. Initial ChIP-seq peaks were called using MACS v. 2.1.1 (Feng and Liu, 2011) with permissive 0.7 q-value cut-off and fragment size of 150bp against summed ChIP-seq input. To generate broad peak calls, we used the modified IDR procedure (https://www.encodeproject.org/software/idr/) with an IDR threshold of 0.05 to combine replicates. This procedure produced broad peaks which we term “IDR peaks”. The pipeline for generating IDR peaks is available here: https://github.com/Przemol/biokludge/blob/master/macs2_idr/macs2_idr.ipy

To generate sharp peak calls set, the IDR calls were further refined using an adhoc post-processing step, as visually distinct peaks close to each other were sometimes identified as a single peak by MACS2. We identified concave regions within MACS2 IDR peaks using the smoothed second derivative of the mapq0 pileup coverage signal with 250bp kernel (https://github.com/Przemol/biokludge/blob/master/macs2_idr/concave_regions.py). We empirically found the minimum of the second derivative within a concave region to be a good indicator of a visually compelling peak and used concave regions (within IDR peaks) with a threshold of lower than -500 curvature index. Next, we discarded peaks with MACS2 score lower than 100 and peak width lower than 100bp. The resulting

peaks were filtered against combined ENCODE (http://www.broadinstitute.org/~anshul/projects/worm/blacklist/ce10-blacklist.bed.gz) and in-house blacklist (https://gist.github.com/przemol/8a712a2e840f95237f4a4f322f65bee1). to generate our final sharp peak calls, described later as “concave peaks”’.

6.1.4 RNA-seq differential expression analyses for genes

We built exon model based on Ensembl Gene 77 (Nov 2014) database gene annotation liftedOver to ce10/WS220. Tag counts for each gene were extracted from BAM alignment files using HTSeq method working in union mode and implemented in R. These values were used to build an expression matrix. The differential gene expression between N2 and mutant backgrounds was tested using DESeq2 procedure (Anders & Huber 2010). Reads per kilobase of exon model per million mapped reads (RPKM) normalized expression values were generated using "median ratio method" (Equation 5 in Anders and Huber, 2010). Table containing RPKM values, maximum posterior estimates of log2 FC (LFC) and statistical significance estimates for each gene is available in supplementary material. We used false discovery rate (FDR) < 0.01 and LFC > 1 to call genes up-regulated, and FDR < 0.01 and LFC < -1 to call genes down-regulated. In addition, the oscillating genes were removed from up- and down-regulated lists, as described in Evans at al. (http://biorxiv.org/content/early/2016/07/17/063990).

6.1.5 RNA-seq differential expression analyses for repeats

We built repetitive elements model based on Dfam 2.0 (Sept 2015, http://dfam.org/) database. The model contained 62331 individual repeats divided into 184 families. Since individual repeats did not have unique identifiers (UID), we named them based on genomic position in “chromosome:start-end” convention, e.g. “chrI:10773-11032”. We assessed differential expression based on both individual repeats and families combined

models. Tag counts for each repeat or repeat family were extracted from BAM alignment files using HTSeq method working in union mode and implemented in R. These values were used to build expression matrixes. The differential repeat expression between N2 and mutant backgrounds was tested using DESeq2 procedure (Love, Simon and Huber 2014). Reads per kilobase of repeat model per million mapped reads (RPKM) normalized expression values were generated using "median ratio method" (Equation 5 in Anders and Huber, 2010). Table containing RPKM values, maximum posterior estimates of log2 FC (LFC) and statistical significance estimates for each repeat is available in supplementary material. We used false discovery rate (FDR) < 0.01, and LFC > 0 and LFC < 0 to call repeats up- and down-regulated respectively. In addition we filtered out repeats overlapping with up- or down regulated genes in each mutant background separately. For purpose of filtering the genes were found using the procedure described above with more permissive cutoffs: FDR < 0.05 and no LFC cutoff (< or > 0 to call up- and down-regulated).

To assess if individual repeats are truly expressed we repeated the above procedure, counting only reads with BWA mapping quality over 10. This procedure discarded all multi-mapping reads, constructing expression matrix with tags specific only to given loci, described later as “mapq10”.

6.1.6 Venn diagrams and UpSet plots for peaks summarized by union

The Venn diagram sum to the number of overlapping ranges present in the assayed peak calls. We created the summary peak call super set by reducing overlapping ranges (union operation) for five heterochromatin factors peak calls. We termed this set “Any5”. For each factor Any5 regions overlapping with peak calls were counted. The numbers of overlaps were assigned to corresponding positions on the diagrams. Venn

diagrams were plotted using VennDiagram R package (Hanbo Chen 2016), and UpSet plots were generated as described in Lex and Gehlenborg, Nature Methods, 2014.

6.1.7 Telomere enrichment

Telomere enrichment for ChIP-seq factors were determined by counting reads with repetitive “GCCTAA” motif. All reads were extracted from BAM files (including not aligned ones) and trimmed to 36bp - the shortest read length of ChIP-seq experiments. Then the number of “GCCTAA” motifs was counted for each read using Biostrings R package. To determine the background levels of motif enrichment 129 input experiments were subjected to the same procedure. The depth of reads coverage normalised fold enrichment over mean of inputs experiments was calculated for following count bins - zero or 1 motif, 2 - 4 motifs and more than 5 motifs. Having 5 or 6 “GCCTAA” motifs in 36bp read strongly identifies this read as telomeric, so we further analysed fold enrichment for this bin. To assess the statistical significance of enrichment we used one sided Mann–Whitney U test (2 replicates for each factor vs. input background of 129 experiments) and reported the p-values.

6.1.8 Mean signal distribution plots and heatmaps

The summarized signal plots and heatmaps for histone marks and DNA associated factors were created using SeqPlots exploratory analyses and plotting tool (Stempor & Ahringer 2016).

6.1.9 Venn diagrams and up-set plots

The Venn diagram sums to the total number of ranges present in the assayed peak calls. We created the summary peak call by piling up all peaks and then extracted the peaks overlapping with one or more peak calls. The numbers of overlaps were assigned to corresponding positions on the diagram.

6.1.10 Venn diagrams and UpSet plots for peaks summarized by union

The Venn diagram sums to the number of overlapping ranges present in the assayed peak calls (the loci covered by at least one factor peak call). We created the summary peak call super set by reducing overlapping ranges (union operation) for five heterochromatin factors peak calls. We termed this set “Any5”. For each factor Any5 regions overlapping with peak calls were counted. The numbers of overlaps were assigned to corresponding positions on the diagrams. Venn diagrams were plotted using VennDiagram R package (Hanbo Chen 2016), and UpSet plots were generated as described in Lex and Gehlenborg, Nature Methods, 2014

6.1.11 Assessing piRNA abundance

The following published small RNA datasets were used: *prg-1* (GSM708661), and *hpl-2* (GSM950181) and matching wild type (N2) experiments (GSM708660, GSM950180). The small RNA (sRNA) sequences were acquired from processed alignments. The reads were uniqued, keeping only a single species of sRNA of given sequence. Next reads were subsampled to match the smallest number of unique dataset (530039 unique reads in *prg*-1, GSM708661). Plots in **Figure 117**, **Figure 118**, **Figure 119** and **Figure 120** show that the global population of small RNAs is mostly unaffected in mutant backgrounds. It illustrates that mapping to known sites is necessary to profile functional 21Us and 22G. These samples were filtered only for reads perfectly matching piRNA annotation. The count of piRNAs in each experiment is reported.

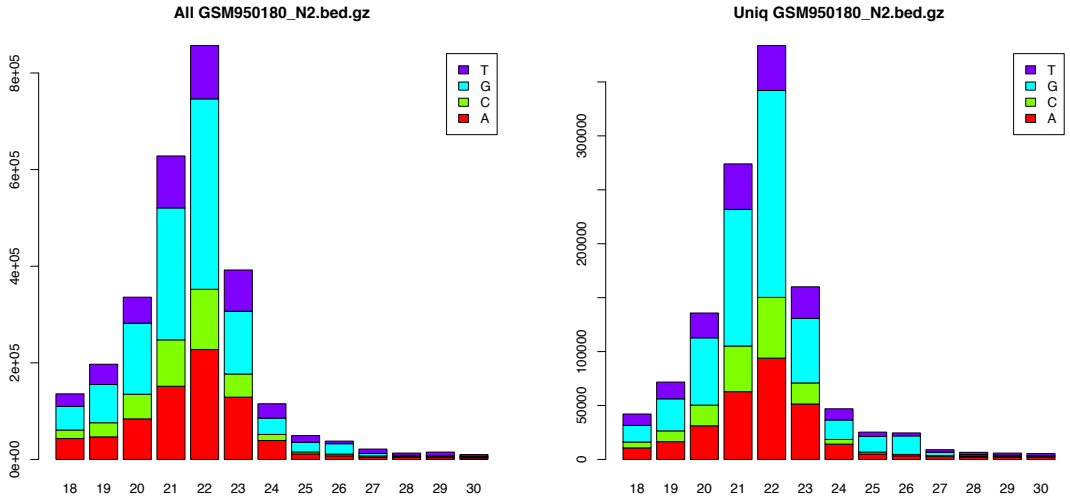


Figure 117 Small RNA abundance in WT sample matching *hpl-2* (GSM950180), read range from 18bp to 30bp. Left panel - all reads, right panel - unique reads. Colours denote the 1st base in the read. The highest population corresponds to 22G.

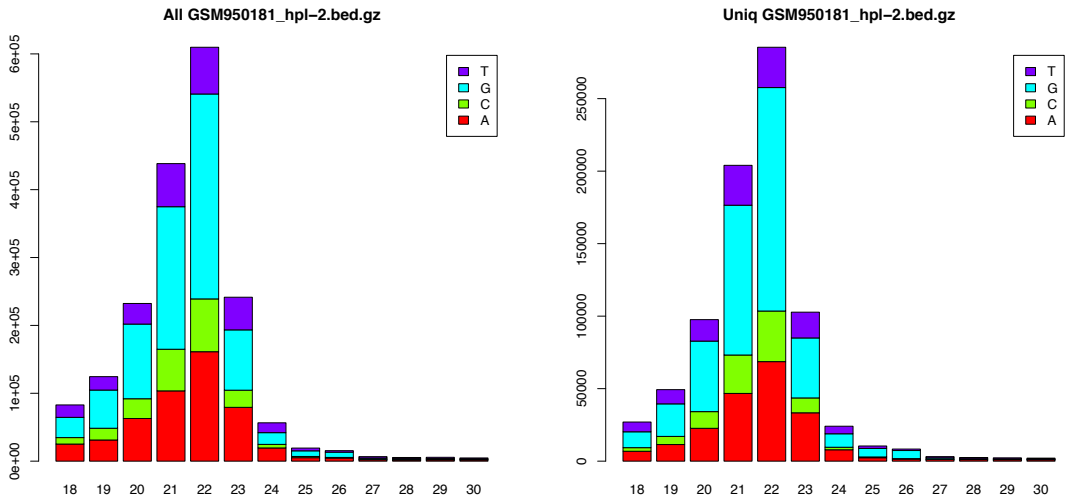


Figure 118 Small RNA abundance in *hpl-2* (GSM950181) sample, read range from 18bp to 30bp. Left panel - all reads, right panel - unique reads. Colours denote the 1st base in the read. The highest population corresponds to 22G.

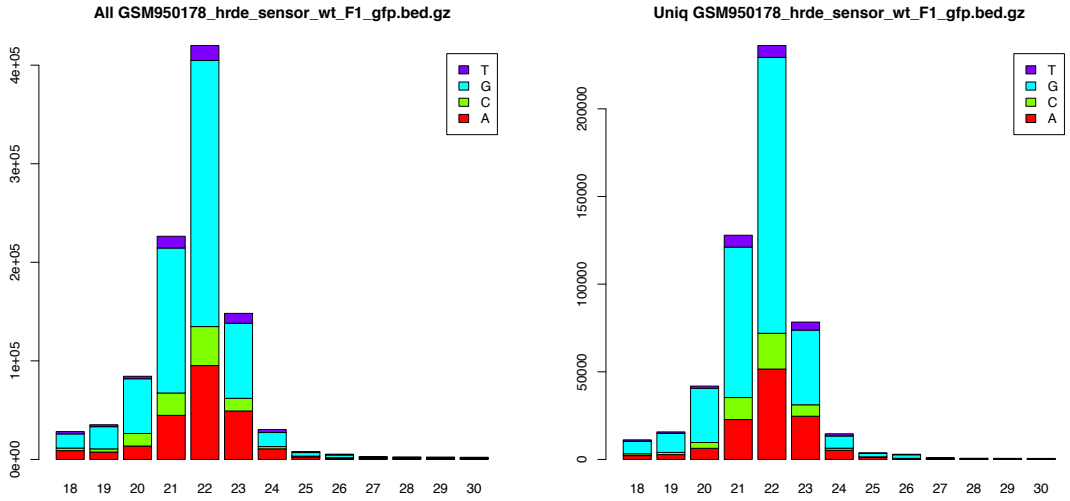


Figure 119 Small RNA abundance in WT matching *nrde-2* (GSM950178), read range from 18bp to 30bp. Left panel - all reads, right panel - unique reads. Colours denote the 1st base in the read. The highest population corresponds to 22G.

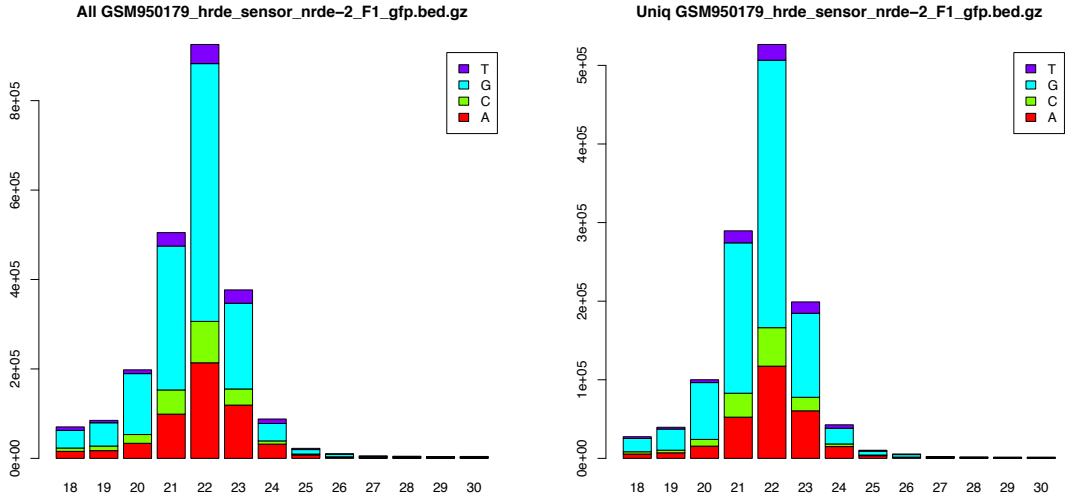


Figure 120 Small RNA abundance in *nrde-2* (GSM950179), read range from 18bp to 30bp. Left panel - all reads, right panel - unique reads. Colours denote the 1st base in the read. The highest population corresponds to 22G.

6.1.12 piRNA annotation

Known piRNA annotation was produced using scoring software from Bartel lab (http://bartellab.wi.mit.edu/software.html), which takes upstream region of piRNA candidates as input and clarifies the region based on Ruby motif score. The annotation we used is available here as the FASTA formatted file: https://gist.githubusercontent.com/Przemol/8dcab5f02587e77d8d320cc6390fc16f/raw/cab3edfea026f528110f8d3a07ab02a3b4ef8008/All_piRNAs_new_reference.fasta. The header is the name according to Batista et al. (type 1) or CAP-seq annotation (type 2) followed by motif score. The sequence represents the piRNA.

6.1.13 piRNA targets mapping

All piRNAs from piRNA annotation (27884) were mapped to *C. elegans* (ce10) genome with up to 4 mismatching bases generating ~10M potential target sites. Next, for each hit the consensus strings between piRNA and reference genome were calculated and recorded using IUPAC encoding. Initial hits were queried based on consensus strings generating final lists of targets. Two different criteria were used - more lax criteria from Lee et al. requiring a perfect match, with no more than one G:U pair, in the seed region (nt 2–8) and allowing up to two mismatches and an additional G:U pair outside of the seed region. This yielded 419708 total sites, and 391173 excluding piRNA self-hits. The more stringent criteria required no more than 2 mismatches between piRNA and target and excluded piRNA self-hits. This yielded 48932 total targets.

6.1.14 Mapping to 22G RNA to piRNA targets

50 bases upstream and 50 based downstream (100bp total) from piRNA target midpoint were considered as possible region generating 22G RNA originating from piRNA. Sequences in these regions were acquired. All 22G RNA from experiments, defined as

reads starting with G and 22nt long, were tested against these sequences. Reads aligning to target sequence with [0 or 1] mismatches were considered as 22G RNA of putative piRNA origin.

6.1.15 Counting 22G RNA targeting repeats

The following small RNA datasets from Ashe et al. (2012) were used: *prg-1* (GSM708661), WT matching *prg-1* (GSM708660), *hpl-2* (GSM950181), WT matching *hpl-2* (GSM950180), *nrde-2* (GSM950179), WT matching *nrde-2* (GSM950178). Uniquely matching positions in each dataset were determined and the smallest number (530039, in the *prg-1* dataset) subsampled from each. piRNA number was then determined by calculating the number matching the piRNAs annotated in Batista et al. (2008) or Weick and Miska (2014) (n = 27884 piRNAs). piRNA targets were determined as in Lee et al. (2012):

- required a perfect match, with no more than one G:U pair, in the seed region (nucleotide 2–8)
- required up to two mismatches and an additional G:U pair outside of the seed region

These criteria yielded a map of 419708 total sites (out of 10M regional sites with more lax criteria, allowing 4 miss matches), and 391173 when excluding self-hits. piRNA dependent 22Gs were also defined as in Lee et al. (2012), as 22G RNAs that mapped in 100 bp windows centred at piRNA target sites, allowing zero or one mismatch.

6.2 Methods used in promoters and open chromatin study

6.2.1 External data sets

C. elegans ChIP-seq data H3K4me3 (modENCODE_5166) and H3K4me1 (modENCODE_5158) and Drosophila ChIP-seq data H3K4me3 (modENCODE_789) and H3K4me1 (modENCODE_777) were obtained from modENCODE (http://www.modencode.org/). MNase digested mononucleosome data for *C. elegans* embryos (GSM468574) (Ooi et al. 2010) and human K562 cells (GSM920557) (The ENCODE Project Consortium 2012) were downloaded from the Gene Expression Omnibus. H3K4me3 (wgEncodeBroadHistoneHepg2H3k4me3StdSig.bigWig) and H3K4me1 (wgEncodeBroadHistoneHepg2H3k04me1StdSig.bigWig) ChIP-seq data in HepG2 cells were obtained from the ENCODE Project (The ENCODE Project Consortium 2012; AP Boyle, CL Araya, C Brdlik, P Cayting, C Cheng, Y Cheng, K Gardner, L Hillier, J Janette, L Jiang, et al., in prep) (http://genome.ucsc.edu/ENCODE/downloads.html). TF ChIP-seq data sets used in the HOT region study can be downloaded from http://anshul.kundaje.net/projects/modencode (for *C. elegans* and Drosophila) and http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/ (for human).

6.2.2 Defining the overlaps of transcription factor binding sites

We used modENCODE and ENCODE collections of TF mapping data for 90 *C. elegans* factors, 54 *D. melanogaster* factors, and 159 human factors (The ENCODE Project Consortium 2007, 2012; Gerstein et al. 2010, 2012; The modENCODE Project Consortium 2010; Nègre et al. 2011; Niu et al. 2011; AP Boyle, CL Araya, C Brdlik, P Cayting, C Cheng, Y Cheng, K Gardner, L Hillier, J Janette, L Jiang, et al., in prep). Data sets for a given factor were merged into single files to prevent double counting of

factors; then at every base, the number of factors where a peak was present was counted. The occupancy of each region is the number of unique factors found inside. For each region of TF overlap, the "core region" was defined as the range with local maxima of peak call coverage. This process identified 35,062 *C. elegans* regions bound by 1–87 factors, 32,168 *D. melanogaster* by 1–37 factors, and 73,7151 human regions bound by 1–138 factors. For all analyses, we defined HOT regions as those in the top 1% of occupancy (376 for *C. elegans*, 341 for *D. melanogaster*, and 7419 for humans) and "COLD" regions as single factor binding regions (13,425 for *C. elegans*, 18,177 for *D. melanogaster*, and 314,323 for humans). The following genome versions were used: hg19 for human, ce10/WS220 for *C. elegans*, and dm3 for Drosophila.

HOT and COLD core regions were scored as overlapping a promoter defined by ±500 bp of a transcript start site. Coding transcript start sites were downloaded from Ensembl genes 71 (human (GRCh37.p10) and *D. melanogaster* (BDGP5). For *C. elegans*, we collected and pooled all coding TSSs recently identified based on capped RNA sequencing (Chen et al. 2013; Kruesi et al. 2013).

6.2.3 Calculation of CpG density

CpG density, GC content, and observed over expected CpG ratio tracks were calculated in sliding windows of 200-bp size and 1-bp shift, reporting calculated values in the middle of the corresponding range. CpG promoter content was calculated at −200 bp to *C. elegans* TSSs and ±500 bp of human TSSs. These calculated CpG values for all protein coding promoters—n = 10,106 for *C. elegans* from Chen et al. (2013) and n = 129,604 for human from Ensembl GRCh37.p12—were ranked to extract promoters harbouring the indicated percentile of CpG content.

6.2.4 Gene expression data

Ubiquitously expressed genes in *C. elegans* were defined as those showing expression (FDR <0.05) in embryonic gut, pan neuron, body wall muscle, germline, and hypodermis in tissue-specific RNA-seq profiling data sets (Spencer et al. 2011). Human ubiquitously expressed genes were defined as those detectably expressed in all examined cell lines (n = 34) in human RNA-seq data sets from the ENCODE Project (The ENCODE Project Consortium 2011) data from http://genome.crg.es/encode_RNA_dashboard/hg19/. Genes with high CpG promoters were those having a promoter in the top 20% CpG band but no promoter in the bottom 80% CpG band (n = 2215 for *C. elegans* and n = 7710 for human). Genes with low CpG promoters were those having a promoter in the bottom 20% CpG band but no promoter in the top 80% CpG band (n = 1764 for *C. elegans* and n = 1834 for human genes).

6.2.5 ChIP-seq profiles, rBEADS normalisation and peak calls

CFP-1 and H3K4me3 ChIP-seq data sets were aligned using BWA with default settings (Li and Durbin 2009), normalized using BEADS (Cheung et al. 2011), then converted to ratios of BEADS scores (enrichment relative to input) and then Z-scored. Peaks were called using MACS v. 2.0.10 software (Feng et al. 2011) with 1×10^{-10} P-value cutoff. Peak calls from each replicate were intersected, and regions present in both were kept. The IGV Genome Browser was applied for visualization (Robinson et al. 2011).

6.2.6 Differential ChIP-seq signal over gene bodies

We extracted gene annotation from the Ensembl Gene 74 (December 2013) database using Biomart. The annotation was liftedOver to ce10/WS220 reference genome. Transcripts were divided into promoter region - +/- 500 bp of annotated transcript start sites (TSS) and gene bodies - TSS + 500 bp to annotated transcript termination sites

(TTS). The average HTZ-1 ChIP-seq signal for gene bodies regions was extracted from reads coverage and BEADS normalized tracks using bigWigSummary utility from UCSC user tools (Kent et al.). These values were used to perform test for differential gene body signal between wild-type (N2) and *lin-35* mutant background replicates using DeSeq2 procedure (Anders and Huber, 2010). Finally, I produced tables containing mean signals, log2 fold change (FC) between wild type and mutant samples and statistical significance estimate (p-value).

6.3 Nonparametric Sparse Factor Analyses (NSFA) R implementation and testing

In order to apply NSFA method to ChIP-seq data I had to devise a more efficient implementation of NSFA to handle much higher dimensionality of data – initial implementation was tested on limited gene expression dataset with hundreds of data points, while binned ChIP-seq experiments have hundred thousand to millions of data points. I decided to implement the whole algorithm *de novo* in R language (interface), and actual computations in C++ language using Armadillo C++ library for linear algebra & scientific computing. After making this implementation work I wanted to ensure numerical consistency between NSFA reference implementation (MATLAB) and new implementation (R/C++). This problem turned out to be non-trivial to solve. Gibbs sampling method that is in core of NSFA algorithm relies heavy on random number generation. R and MATLAB, though theoretically both implement the same pseudo-random number generators, have a very different actual implementations and generate non-consistent sequences from same initial state. For this reason, I had to find a way for R to employ the random number generator built into MATLAB.

6.3.1 Reference implementation and test datasets

(1) The reference implementation of NSFA in MATLAB was prepared by David Knowles, as described in https://projecteuclid.org/euclid.aoas/1310562732 and Chapters 2-3 of his thesis (http://cs.stanford.edu/people/davidknowles/daknowles_thesis.pdf)

(2) I implemented the same algorithm in R/C++. The MATLAB implementation is not efficient enough to work with the dimensionality of data used in genomics. Also, the implementation in closed, commercial tool inhibits the efficient cluster deployment (Matlab's per computing core pricing policy generates high costs) and sharing with scientific community.

(3) Test dataset - Breast cancer dataset from West et al. [2007], used in original NSFA paper

6.3.2 Numerical consistency and pseudorandom number generators

NSFA is probabilistic algorithm, heavily relying on Markov Chain Monte Carlo (MCMC) sampling method - the Gibbs sampler. During algorithm iterations the random number generator is called multiple times. The consistent random number generation between R and MATLAB is necessary to ensure numerical consistency between implementations. I considered three approaches to achieve this goal:

- Using the same random number generator with same state to generate consistent numbers
- Call one language from within the other
- Pre-calculate random numbers and store them in text file for future use

The further complication comes from the fact that NSFA (including the initialization and test data preparation stages) requires to generate random numbers from 6 probability distributions:

- Uniform
- Poisson
- Normal
- Gamma
- Binomial
- Multivariate normal (optional, depending on noise parameters generation settings)

6.3.3 Random number generators for uniform distribution

Both MATLAB and R use Mersenne Twister (MT) to generate random numbers (http://dl.acm.org/citation.cfm?doid=272991.272995). However, the initiation of the Mersenne Twister's state from seed number differs between the languages. MT19937 implementation in MATLAB uses integer values to save internal state. The state in R is coded as 626 signed integer vectors with 1st number equals to constant 403, 2nd denoting the position (increased by one after each pass of algorithm) and following 624 encoding the actual state. In MATLAB a 625 element, 32-bit unsigned integer column vector is used, where 624 elements denote the state and last one position. Despite this significant difference the vector generated in one language, e.g. MATLAB can be saved and mapped to other language, e.g. R. The final difference in random number generation comes from the precision - MATLAB generates the numbers in double precision, which requires 2 passes over MT, while R generates numbers in single precision - only one pass over MT. This means that either the state will have to be re-established after each generation or MTs in R and MATLAB will run out of sync just after one number is generated. Fortunately, MATLAB can be set up to generate single precision numbers, making uniform distribution numbers in MATLAB and R reproducible over millions of iterations. In conclusion, MATLAB and R can consistently produce same random numbers from uniform distribution.

6.3.4 Random number generators for non-uniform distributions

Poisson, normal, gamma, binomial, and other distribution's random numbers are generated by taking sample from uniform distribution and transforming it by appropriate function/algorithm. The exact implementation of generating pseudorandom numbers from non-uniform distributions differs between R and MATLAB - for example for Gaussian distribution the number of over MT differs, which again means the MTs will run out of sync after one Gaussian number is generated. Re-implementing all

required functions either in R or MATLAB is non-trivial and time-consuming task. The solution for this problem is R.matlab package, which allows MATLAB code execution from within R. This is not an optimal solution performance-wise, but for small problems it is sufficient. The final step involved writing the wrapper functions, that will replace original R pseudorandom functions with calls to MATLAB. In conclusion, for non-uniform distributions it is easier to call MATLAB from R in order to produce consistent, high quality pseudorandom numbers in both environments.

6.3.5 C++/R and MATLAB implementations of NSFA are numerically consistent

After applying the method for getting consistent random numbers in R and MATLAB I tested the numerical consistency between these two methods using small expression datasets - Breast cancer dataset from (West *et al.* 2007). With consistent random number generators, I reproduced the same result in both environments. In conclusion, MATLAB and R implementations of NSFA are numerically consistent.

6.4 Weblink to software and resources

MACS2 - https://github.com/taoliu/MACS

BWA - http://bio-bwa.sourceforge.net/

Samtools - http://www.htslib.org/

R - https://www.r-project.org/

Bioconductor - http://bioconductor.org/

IDR2 - https://github.com/ENCODE-DCC/chip-seq-pipeline/tree/master/dnanexus/idr2

Ensembl - https://www.ensembl.org

Dfam 2.0 - http://dfam.org/

7 REFERENCES

Ahringer, J. (2000). NuRD and SIN3: Histone deacetylase complexes in development. *Trends in Genetics*, **16**(1997), 351–356.

Alló, M., & Kornblihtt, A. R. (2010). Gene Silencing: Small RNAs Control RNA Polymerase II Elongation. *Current Biology*, **20**(17), R704–R707.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106.

Ashe, A., Sapetschnig, A., Weick, E. M., Mitchell, J., Bagijn, M. P., Cording, A. C., Doebley, A. L., Goldstein, L. D., Lehrbach, N. J., Le Pen, J., Pintacuda, G., Sakaguchi, A., Sarkies, P., Ahmed, S., & Miska, E. A. (2012). piRNAs can trigger a multigenerational epigenetic memory in the germline of C. elegans. *Cell*, **150**(1), 88–99.

Batista, P. J., Ruby, J. G., Claycomb, J. M., Chiang, R., Fahlgren, N., Kasschau, K. D., Chaves, D. A., Gu, W., Vasale, J. J., Duan, S., Conte, D., Luo, S., Schroth, G. P., Carrington, J. C., Bartel, D. P., & Mello, C. C. (2008). PRG-1 and 21U-RNAs Interact to Form the piRNA Complex Required for Fertility in C. elegans. *Molecular Cell*, **31**(1), 67–78.

Bauer, D. F. (1972). Constructing Confidence Sets Using Rank Statistics. *Journal of the American Statistical Association*, **67**(339), 687–690.

Beerman, I., Seita, J., Inlay, M. A., Weissman, I. L., & Rossi, D. J. (2014). Quiescent hematopoietic stem cells accumulate DNA damage during aging that is repaired upon entry into cell cycle. *Cell Stem Cell*, **15**(1), 37–50.

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome.

Nature, **489**(7414), 57–74.

Bian, C., Xu, C., Ruan, J., Lee, K. K., Burke, T. L., Tempel, W., Barsyte, D., Li, J., Wu, M., Zhou, B. O., Fleharty, B. E., Paulson, A., Allali-Hassani, A., Zhou, J.-Q., Mer, G., Grant, P. A., Workman, J. L., Zang, J., & Min, J. (2011). SGF29 BINDS HISTONE H3K4ME2/3 AND IS REQUIRED FOR SAGA COMPLEX RECRUITMENT AND HISTONE H3 ACETYLATION. *The EMBO Journal*, **30**(14), 2829–42.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C.

N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X. X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H. H., Ng, P., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X. X., Xu, M., Haidar, J. N. S., Yu, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R.

W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., de Jong, P. J., Stamatoyannopoulos, J. A., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Dutta, A., Guigó, R., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Flicek, P., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X. X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Dermitzakis, E. T., Margulies, E. H., Hubbard, T., Myers, R. M., Rogers, J.,

Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Snyder, M., Birney, E., Struhl, K., Gerstein, M., Antonarakis, S. E., Gingeras, T. R., Brown, J. B., Flicek, P., Fu, Y., Keefe, D., Birney, E., Denoeud, F., Gerstein, M., Green, E. D., Kapranov, P., Karaöz, U., Myers, R. M., Noble, W. S., Reymond, A., Rozowsky, J., Struhl, K., Siepel, A., Stamatoyannopoulos, J. A., Taylor, C. M., Taylor, J., Thurman, R. E., Tullius, T. D., Washietl, S., Zheng, D., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Collins, F. S., Margulies, E. H., Cooper, G. M., Asimenos, G., Thomas, D. J., Dewey, C. N., Siepel, A., Birney, E., Keefe, D., Hou, M., Taylor, J., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Brown, J. B., Huang, H., Zhang, N. R., Bickel, P., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Gerstein, M., Antonarakis, S. E., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Pachter, L., Green, E. D., Sidow, A., Weng, Z., Trinklein, N. D., Fu, Y., Zhang, Z. D., Karaöz, U., Barrera, L., Stuart, R., Zheng, D., Ghosh, S., Flicek, P., King, D. C., Taylor, J., Ameur, A., Enroth, S., Bieda, M. C., Koch, C. M., Hirsch, H. A., Wei, C.-L., Cheng, J., Kim, J., Bhinge, A. A., Giresi, P. G., Jiang, N., Liu, J., Yao, F., Sung, W.-K., Chiu, K. P., Vega, V. B., Lee, C. W. H. H., Ng, P., Shahab, A., Sekinger, E. A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Clelland, G. K., Wilcox, S., Dillon, S. C., Andrews, R. M., Fowler, J. C., Couttet, P., James, K. D., Lefebvre, G. C., Bruce, A. W., Dovey, O. M., Ellis, P. D., Dhami, P., Langford, C. F., Carter, N. P., Vetrie, D., Kapranov, P., Nix, D. A., Bell, I., Patel, S., Rozowsky, J., Euskirchen, G., Hartman, S., Lian, J., Wu, J., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Hoon Kim, T., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X. X., Xu, M., Haidar,

J. N. S., Yu, Y., Birney*, E., Weissman, S., Ruan, Y., Lieb, J. D., Iyer, V. R., Green, R. D., Gingeras, T. R., Wadelius, C., Dunham, I., Struhl, K., Hardison, R. C., Gerstein, M., Farnham, P. J., Myers, R. M., Ren, B., Snyder, M., Thomas, D. J., Rosenbloom, K., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Haussler, D., Kent, W. J., Dermitzakis, E. T., Armengol, L., Bird, C. P., Clark, T. G., Cooper, G. M., de Bakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Thomas, D. J., Woodroffe, A., Batzoglou, S., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Hardison, R. C., Huppert, J., Sidow, A., Taylor, J., Trumbower, H., Zody, M. C., Guigó, R., Mullikin, J. C., Abecasis, G. R., Estivill, X., Birney, E., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., & de Jong, P. J. (2007). IDENTIFICATION AND ANALYSIS OF FUNCTIONAL ELEMENTS IN 1% OF THE HUMAN GENOME BY THE ENCODE PILOT PROJECT. *Nature*, **447**(7146), 799–816.

Boeck, M. E., Huynh, C., Gevirtzman, L., Thompson, O. A., Wang, G., Kasper, D. M., Reinke, V., Hillier, L. W., & Waterston, R. H. (2016). THE TIME-RESOLVED TRANSCRIPTOME OF C. ELEGANS. *Genome Research*, **26**(10), 1441–1450.

Boyle, A. P., Araya, C. L., Brdlik, C., Cayting, P., Cheng, C., Cheng, Y., Gardner, K., Hillier, L. W., Janette, J., Jiang, L., Kasper, D., Kawli, T., Kheradpour, P., Kundaje, A., Li, J. J., Ma, L., Niu, W., Rehm, E. J., Rozowsky, J., Slattery, M., Spokony, R., Terrell, R., Vafeados, D., Wang, D., Weisdepp, P., Wu, Y.-C., Xie, D., Yan, K.-K., Feingold, E. A., Good, P. J., Pazin, M. J., Huang, H., Bickel, P. J.,

Brenner, S. E., Reinke, V., Waterston, R. H., Gerstein, M., White, K. P., Kellis, M., & Snyder, M. (2014). COMPARATIVE ANALYSIS OF REGULATORY INFORMATION AND CIRCUITS ACROSS DISTANT SPECIES. *Nature*, **512**(7515), 453–456.

Britten, R. J. (1995). ACTIVE GYPSY/TY3 RETROTRANSPOSONS OR RETROVIRUSES IN CAENORHABDITIS ELEGANS. *Proceedings of the National Academy of Sciences of the United States of America*, **92**(2), 599–601.

Brivanlou, A. H., & Darnell, J. E. (2002). SIGNAL TRANSDUCTION AND THE CONTROL OF GENE EXPRESSION. *Science (New York, N.Y.)*, **295**(5556), 813–8.

Brown, D. A., Cerbo, V. Di, Feldmann, A., Kutateladze, T. G., Koseki, H., Klose Correspondence, R. J., Ahn, J., Ito, S., Blackledge, N. P., Nakayama, M., Mcclellan, M., Dimitrova, E., Turberfield, A. H., Long, H. K., King, H. W., Kriaucionis, S., Schermelleh, L., & Klose, R. J. (2017a). THE SET1 COMPLEX SELECTS ACTIVELY TRANSCRIBED TARGET GENES VIA MULTIVALENT INTERACTION WITH CPG ISLAND CHROMATIN. *CellReports*, **20**, 2313–2327.

Brown, D. A., Di Cerbo, V., Feldmann, A., Ahn, J., Ito, S., Blackledge, N. P., Nakayama, M., McClellan, M., Dimitrova, E., Turberfield, A. H., Long, H. K., King, H. W., Kriaucionis, S., Schermelleh, L., Kutateladze, T. G., Koseki, H., & Klose, R. J. (2017b). THE SET1 COMPLEX SELECTS ACTIVELY TRANSCRIBED TARGET GENES VIA MULTIVALENT INTERACTION WITH CPG ISLAND CHROMATIN. *Cell Reports*, **20**(10), 2313–2327.

Brown, T. A. (Terence A. . (1998). *Molecular biology labfax*, Academic Press. Retrieved from https://books.google.co.uk/books/about/Molecular_Biology_Labfax.html?id=02Z1ZAtmGBIC&redir_esc=y

Buckley, B. A., Burkhart, K. B., Gu, S. G., Spracklin, G., Kershner, A., Fritz, H., Kimble, J., Fire, A., & Kennedy, S. (2012). A NUCLEAR ARGONAUTE PROMOTES MULTIGENERATIONAL EPIGENETIC INHERITANCE AND GERMLINE IMMORTALITY.

Nature, **489**(7416), 447–451.

Burkhart, K. B., Guang, S., Buckley, B. A., Wong, L., Bochner, A. F., & Kennedy, S. (2011). A PRE-MRNA–ASSOCIATING FACTOR LINKS ENDOGENOUS SIRNAS TO CHROMATIN REGULATION. *PLoS Genetics*, **7**(8), e1002249.

Ceol, C. J., Stegmeier, F., Harrison, M. M., & Horvitz, H. R. (2006). IDENTIFICATION AND CLASSIFICATION OF GENES THAT ACT ANTAGONISTICALLY TO LET-60 RAS SIGNALING IN CAENORHABDITIS ELEGANS VULVAL DEVELOPMENT. *Genetics*, **173**(2), 709–726.

Checchi, P. M., & Engebrecht, J. (2011). CAENORHABDITIS ELEGANS HISTONE METHYLTRANSFERASE MET-2 SHIELDS THE MALE X CHROMOSOME FROM CHECKPOINT MACHINERY AND MEDIATES MEIOTIC SEX CHROMOSOME INACTIVATION. *PLoS Genetics*, **7**(9), e1002267.

Chen, R. A.-J. J., Down, T. a., Stempor, P., Chen, Q. B., Egelhofer, T. a., Hillier, L. W., Jeffers, T. E., & Ahringer, J. (2013). THE LANDSCAPE OF RNA POLYMERASE II TRANSCRIPTION INITIATION IN C. ELEGANS REVEALS PROMOTER AND ENHANCER ARCHITECTURES. *Genome Research*, **23**(8), 1339–47.

Chen, R. A.-J., Stempor, P., Down, T. a., Zeiser, E., Feuer, S. K., & Ahringer, J. (2014a). EXTREME HOT REGIONS ARE CPG DENSE PROMOTERS IN C. ELEGANS AND HUMANS. *Genome Research*. doi:10.1101/gr.161992.113

Chen, R. A.-J., Stempor, P., Down, T. A., Zeiser, E., Feuer, S. K., & Ahringer, J. (2014b). EXTREME HOT REGIONS ARE CPG-DENSE PROMOTERS IN C. ELEGANS AND HUMANS. *Genome Research*, **24**(7), 1138–46.

Clouaire, T., Webb, S., & Bird, A. (2014). CFP1 IS REQUIRED FOR GENE EXPRESSION-DEPENDENT H3K4 TRIMETHYLATION AND H3K9 ACETYLATION IN EMBRYONIC STEM CELLS. *Genome Biology*, **15**(9), 451.

Corà, D., Di Cunto, F., Caselle, M., & Provero, P. (2007). IDENTIFICATION OF CANDIDATE REGULATORY SEQUENCES IN MAMMALIAN 3' UTRS BY STATISTICAL

ANALYSIS OF OLIGONUCLEOTIDE DISTRIBUTIONS. *BMC Bioinformatics*, **8**, 174.

Coustham, V., Bedet, C., Monier, K., Schott, S., Karali, M., & Palladino, F. (2006). THE C. ELEGANS HP1 HOMOLOGUE HPL-2 AND THE LIN-13 ZINC FINGER PROTEIN FORM A COMPLEX IMPLICATED IN VULVAL DEVELOPMENT. *Developmental Biology*, **297**(2), 308–322.

Crick, F. (1970). CENTRAL DOGMA OF MOLECULAR BIOLOGY. *Nature*, **227**(5258), 561–3.

Cui, M., Chen, J., Myers, T. R., Hwang, B. J., Sternberg, P. W., Greenwald, I., & Han, M. (2006). SYNMUV GENES REDUNDANTLY INHIBIT LIN-3/EGF EXPRESSION TO PREVENT INAPPROPRIATE VULVAL INDUCTION IN C. ELEGANS. *Developmental Cell*, **10**(5), 667–672.

Das, P. P., Bagijn, M. P., Goldstein, L. D., Woolford, J. R., Lehrbach, N. J., Sapetschnig, A., Buhecha, H. R., Gilchrist, M. J., Howe, K. L., Stark, R., Matthews, N., Berezikov, E., Ketting, R. F., Tavaré, S., & Miska, E. A. (2008). PIWI AND PIRNAS ACT UPSTREAM OF AN ENDOGENOUS SIRNA PATHWAY TO SUPPRESS TC3 TRANSPOSON MOBILITY IN THE CAENORHABDITIS ELEGANS GERMLINE. *Molecular Cell*, **31**(1), 79–90.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). TOPOLOGICAL DOMAINS IN MAMMALIAN GENOMES IDENTIFIED BY ANALYSIS OF CHROMATIN INTERACTIONS. *Nature*, **485**(7398), 376–80.

Ecco, G., Imbeault, M., & Trono, D. (2017). KRAB ZINC FINGER PROTEINS. *Development*, **144**(15), 2719–2729.

Eissenberg, J. C. (2001). MOLECULAR BIOLOGY OF THE CHROMO DOMAIN: AN ANCIENT CHROMATIN MODULE COMES OF AGE. *Gene*, **275**(1), 19–29.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). QGRAPH : NETWORK VISUALIZATIONS OF RELATIONSHIPS IN PSYCHOMETRIC DATA. *Journal of Statistical Software*, **48**(4). doi:10.18637/jss.v048.i04

Eskeland, R., Eberharter, A., & Imhof, A. (2007). HP1 BINDING TO CHROMATIN

METHYLATED AT H3K9 IS ENHANCED BY AUXILIARY FACTORS. *Molecular and Cellular Biology*, **27**(2), 453–65.

Fay, D. S., & Yochem, J. (2007). THE SYNMUV GENES OF CAENORHABDITIS ELEGANS IN VULVAL DEVELOPMENT AND BEYOND. *Developmental Biology*, **306**(1), 1–9.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J. Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., & Andrau, J.-C. (2012). CPG ISLANDS AND GC CONTENT DICTATE NUCLEOSOME DEPLETION IN A TRANSCRIPTION-INDEPENDENT MANNER AT MAMMALIAN PROMOTERS. *Genome Research*, **22**(12), 2399–2408.

Foygel, R., & Drton, M. (2010). EXTENDED BAYESIAN INFORMATION CRITERIA FOR GAUSSIAN GRAPHICAL MODELS. Retrieved from http://arxiv.org/abs/1011.6640

Friedman, J., Hastie, T., & Tibshirani, R. (2014). GLASSO: GRAPHICAL LASSO-ESTIMATION OF GAUSSIAN GRAPHICAL MODELS.

Fruchterman, T. M. J., & Reingold, E. M. (1991). GRAPH DRAWING BY FORCE-DIRECTED PLACEMENT. *Software: Practice and Experience*, **21**(11), 1129–1164.

Gardiner-Garden, M., & Frommer, M. (1987). CPG ISLANDS IN VERTEBRATE GENOMES. *Journal of Molecular Biology*, **196**(2), 261–82.

Garrigues, J. M., Sidoli, S., Garcia, B. A., & Strome, S. (2015). DEFINING HETEROCHROMATIN IN C. ELEGANS THROUGH GENOME-WIDE ANALYSIS OF THE HETEROCHROMATIN PROTEIN 1 HOMOLOG HPL-2. *Genome Research*, **25**(1), 76–88.

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A.,

Farnham, P. J., Myers, R. M., Weissman, S. M., & Snyder, M. (2012). ARCHITECTURE OF THE HUMAN REGULATORY NETWORK DERIVED FROM ENCODE DATA. *Nature*, **489**(7414), 91–100.

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, a., Ikegami, K., Alves, P., Chateigner, a., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, a., Niu, W., Rhrissorrakrai, K., Agarwal, a., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, a., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, a. F., Desai, a., Dick, L., Dose, a. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. a., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, a., Holster, H., Hyman, T., Iniguez, a. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, a., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecenas, D., Merrihew, G., Miller, D. M., Muroyama, a., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. a., Rajewsky, N., Ratsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, a., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan, K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D., & Waterston, R. H. (2010). INTEGRATIVE ANALYSIS OF THE CAENORHABDITIS ELEGANS GENOME BY THE MODENCODE PROJECT. *Science*, **330**(6012), 1775–1787.

Gerstein, M. B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J. B., Davis, C. A., Hillier, L., Sisu, C., Li, J. J., Pei, B., Harmanci, A. O., Duff, M. O., Djebali, S., Alexander, R. P., Alver, B. H., Auerbach, R., Bell, K., Bickel, P. J., Boeck, M. E., Boley, N. P., Booth, B. W., Cherbas, L., Cherbas, P., Di, C., Dobin, A., Drenkow, J., Ewing, B., Fang, G., Fastuca, M., Feingold, E. A., Frankish, A., Gao, G., Good, P. J., Guigó, R., Hammonds, A., Harrow, J., Hoskins, R. A., Howald, C., Hu, L., Huang, H., Hubbard, T. J. P., Huynh, C., Jha, S., Kasper, D., Kato, M., Kaufman, T. C., Kitchen, R. R., Ladewig, E., Lagarde, J., Lai, E., Leng, J., Lu, Z., MacCoss, M., May, G., McWhirter, R., Merrihew, G., Miller, D. M., Mortazavi, A., Murad, R., Oliver, B., Olson, S., Park, P. J., Pazin, M. J., Perrimon, N., Pervouchine, D., Reinke, V., Reymond, A., Robinson, G., Samsonova, A., Saunders, G. I., Schlesinger, F., Sethi, A., Slack, F. J., Spencer, W. C., Stoiber, M. H., Strasbourger, P., Tanzer, A., Thompson, O. A., Wan, K. H., Wang, G., Wang, H., Watkins, K. L., Wen, J., Wen, K., Xue, C., Yang, L., Yip, K., Zaleski, C., Zhang, Y., Zheng, H., Brenner, S. E., Graveley, B. R., Celniker, S. E., Gingeras, T. R., & Waterston, R. (2014). COMPARATIVE ANALYSIS OF THE TRANSCRIPTOME ACROSS DISTANT SPECIES. *Nature*, **512**(7515), 445–448.

Ghahramani, Z. (2004). UNSUPERVISED LEARNING ∗, 1–32.

Ghahramani, Z. (2011). BAYESIAN NONPARAMETRICS AND THE PROBABILISTIC APPROACH TO MODELLING, 1–27.

Glaser, S., Schaft, J., Lubitz, S., Vintersten, K., van der Hoeven, F., Tufteland, K. R., Aasland, R., Anastassiadis, K., Ang, S. L., & Stewart, A. F. (2006). MULTIPLE EPIGENETIC MAINTENANCE FACTORS IMPLICATED BY THE LOSS OF MLL2 IN MOUSE DEVELOPMENT. *Development*, **133**(8), 1423–32.

Göndör, A., & Ohlsson, R. (2009). CHROMOSOME CROSSTALK IN THREE DIMENSIONS. *Nature*, **461**(September), 212–217.

Grabundzija, I., Messing, S. A., Thomas, J., Cosby, R. L., Bilic, I., Miskey, C., Gogol-

Döring, A., Kapitonov, V., Diem, T., Dalda, A., Jurka, J., Pritham, E. J., Dyda, F., Izsvák, Z., & Ivics, Z. (2016). A HELITRON TRANSPOSON RECONSTRUCTED FROM BATS REVEALS A NOVEL MECHANISM OF GENOME SHUFFLING IN EUKARYOTES. *Nature Communications*, **7**, 10716.

Griffiths, T., & Ghahramani, Z. (2011). THE INDIAN BUFFET PROCESS: AN INTRODUCTION AND REVIEW. *The Journal of Machine Learning Research*, **12**, 1185–1224.

Gu, W., Shirayama, M., Conte, D., Vasale, J., Batista, P. J., Claycomb, J. M., Moresco, J. J., Youngman, E. M., Keys, J., Stoltz, M. J., Chen, C.-C. G., Chaves, D. A., Duan, S., Kasschau, K. D., Fahlgren, N., Yates, J. R., Mitani, S., Carrington, J. C., & Mello, C. C. (2009). DISTINCT ARGONAUTE-MEDIATED 22G-RNA PATHWAYS DIRECT GENOME SURVEILLANCE IN THE C. ELEGANS GERMLINE. *Molecular Cell*, **36**(2), 231–244.

Guang, S., Bochner, A. F., Burkhart, K. B., Burton, N., Pavelec, D. M., & Kennedy, S. (2010). SMALL REGULATORY RNAS INHIBIT RNA POLYMERASE II DURING THE ELONGATION PHASE OF TRANSCRIPTION. *Nature*, **465**(7301), 1097–1101.

Hackett, J. a, Sengupta, R., Zylicz, J. J., Murakami, K., Lee, C., Down, T. a, & Surani, M. A. (2013). GERMLINE DNA DEMETHYLATION DYNAMICS AND IMPRINT ERASURE THROUGH 5-HYDROXYMETHYLCYTOSINE. *Science (New York, N.Y.)*, **339**(6118), 448–52.

Haubold, B., & Wiehe, T. (2006). HOW REPETITIVE ARE GENOMES? *BMC Bioinformatics*, **7**, 541.

He, J., Kallin, E. M., Tsukada, Y., & Zhang, Y. (2008). THE H3K36 DEMETHYLASE JHDM1B/KDM2B REGULATES CELL PROLIFERATION AND SENESCENCE THROUGH P15INK4B. *Nature Structural & Molecular Biology*, **15**(11), 1169–1175.

Higham, N. (2002). COMPUTING THE NEAREST CORRELATION MATRIX - A PROBLEM FROM FINANCE. *IMA Journal of Numerical Analysis 22, 329–343.*

Ho, J. W. K., Jung, Y. L., Liu, T., Alver, B. H., Lee, S., Ikegami, K., Sohn, K., Minoda, A., Tolstorukov, M. Y., Appert, A., Parker, S. C. J., Gu, T., Kundaje, A., Riddle, N. C., Bishop, E., Egelhofer, T. a, Hu, S., Alekseyenko, A. a, Rechtsteiner, A., Asker, D., Belsky, J. a, Bowman, S. K., Chen, Q. B., Chen, R. a, Day, D. S., Dong, Y., Dose, A. C., Duan, X., Epstein, C. B., Ercan, S., Feingold, E. a, Ferrari, F., Garrigues, J. M., Gehlenborg, N., Good, P. J., Haseley, P., He, D., Herrmann, M., Hoffman, M. M., Jeffers, T. E., Kharchenko, P. V, Kolasinska-zwierz, P., Kotwaliwale, C. V, Kumar, N., Langley, S. a, Larschan, E. N., Latorre, I., Libbrecht, M. W., Lin, X., Park, R., Pazin, M. J., Pham, H. N., Plachetka, A., Qin, B., Schwartz, Y. B., Shoresh, N., Stempor, P., Vielle, A., Wang, C., Whittle, C. M., Xue, H., Kingston, R. E., Kim, J. H., Bernstein, B. E., Dernburg, A. F., Pirrotta, V., Kuroda, M. I., Noble, W. S., Tullius, T. D., Kellis, M., Macalpine, D. M., Strome, S., Elgin, S. C. R., & Ad, N. (2014). COMPARATIVE ANALYSIS OFMETAZOAN CHROMATIN ORGANIZATION. *Nature*, **512**(7515), 449–452.

Howe, F. S., Fischl, H., Murray, S. C., & Mellor, J. (2017). IS H3K4ME3 INSTRUCTIVE FOR TRANSCRIPTION ACTIVATION? *BioEssays*, **39**(1), e201600095.

Hu, D., Gao, X., Morgan, M. A., Herz, H. M., Smith, E. R., & Shilatifard, A. (2013a). THE MLL3/MLL4 BRANCHES OF THE COMPASS FAMILY FUNCTION AS MAJOR HISTONE H3K4 MONOMETHYLASES AT ENHANCERS. *Mol Cell Biol*, **33**(23), 4745–54.

Hu, G., Cui, K., Northrup, D., Liu, C., Wang, C., Tang, Q., Ge, K., Levens, D., Crane-Robinson, C., & Zhao, K. (2013b). H2A.Z FACILITATES ACCESS OF ACTIVE AND REPRESSIVE COMPLEXES TO CHROMATIN IN EMBRYONIC STEM CELL SELF-RENEWAL AND DIFFERENTIATION. *Cell Stem Cell*, **12**(2), 180–92.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. a, Lawrence, M., Love, M. I., Macdonald, J., Obenchain, V., Oleś,

A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., & Morgan, M. (2015). ORCHESTRATING HIGH-THROUGHPUT GENOMIC ANALYSIS WITH BIOCONDUCTOR. *Nature Publishing Group*, **12**(2), 115–121.

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J. (2016). THE DFAM DATABASE OF REPETITIVE DNA FAMILIES. *Nucleic Acids Research*, **44**(D1), D81–D89.

Hucks, D. (2008). *TRANSPOSON EXAPTATION IN MAMMALIAN EVOLUTION*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.324.1156&rep=rep1&type=pdf

Illingworth, R. S., & Bird, A. P. (2009). CPG ISLANDS - "A ROUGH GUIDE." *FEBS Letters*, **583**(11), 1713–1720.

Janes, J., Dong, Y., Schoof, M., Serizay, J., Appert, A., Cerrato, C., Woodbury, C., Chen, R., Gemma, C., Huang, N., Kissiov, D., Stempor, P., Steward, A., Zeiser, E., Sauer, S., & Ahringer, J. (2018). CHROMATIN ACCESSIBILITY DYNAMICS ACROSS C. ELEGANS DEVELOPMENT AND AGEING. *BioRxiv*, 279158.

Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P., & Fire, A. Z. (2006). FLEXIBILITY AND CONSTRAINT IN THE NUCLEOSOME CORE LANDSCAPE OF CAENORHABDITIS ELEGANS CHROMATIN. *Genome Research*, **16**, 1505–1516.

Johnson, T. E., & Simpson, V. J. (1985). AGING STUDIES IN C. ELEGANS AND OTHER NEMATODES. In *CRC Handbook of Cell Biology of Aging*, CRC Press, Boca Raton, FL, pp. 481–495.

Kapitonov, V. V, & Jurka, J. (2001). ROLLING-CIRCLE TRANSPOSONS IN EUKARYOTES. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(15), 8714–9.

Käser-Pébernard, S., Müller, F., & Wicky, C. (2014). LET-418/MI2 AND SPR-5/LSD1 COOPERATIVELY PREVENT SOMATIC REPROGRAMMING OF C. ELEGANS GERMLINE

STEM CELLS. *Stem Cell Reports*, **2**(4), 547–559.

Katz, Y., Wang, E. T., Silterra, J., Schwartz, S., Wong, B., Thorvaldsdóttir, H., Robinson, J. T., Mesirov, J. P., Airoldi, E. M., & Burge, C. B. (2015). QUANTITATIVE VISUALIZATION OF ALTERNATIVE EXON EXPRESSION FROM RNA-SEQ DATA. *Bioinformatics*, **31**(14), 2400–2402.

Kellum, R., & Alberts, B. M. (1995). HETEROCHROMATIN PROTEIN 1 IS REQUIRED FOR CORRECT CHROMOSOME SEGREGATION IN DROSOPHILA EMBRYOS. *Journal of Cell Science*, **108 (Pt 4)**, 1419–31.

Kent, W. J., Zweig, a S., Barber, G., Hinrichs, a S., & Karolchik, D. (2010). BIGWIG AND BIGBED: ENABLING BROWSING OF LARGE DISTRIBUTED DATASETS. *Bioinformatics (Oxford, England)*, **26**(17), 2204–7.

Kimura, H., Hayashi-Takanaka, Y., Goto, Y., Takizawa, N., & Nozaki, N. (2008). THE ORGANIZATION OF HISTONE H3 MODIFICATIONS AS REVEALED BY A PANEL OF SPECIFIC MONOCLONAL ANTIBODIES. *Cell Structure and Function*, **33**(1), 61–73.

Kircher, M., Sawyer, S., & Meyer, M. (2012). DOUBLE INDEXING OVERCOMES INACCURACIES IN MULTIPLEX SEQUENCING ON THE ILLUMINA PLATFORM. *Nucleic Acids Research*, **40**(1), e3–e3.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., & Wild, D. L. (2012). BAYESIAN CORRELATED CLUSTERING TO INTEGRATE MULTIPLE DATASETS. *Bioinformatics (Oxford, England)*, **28**(24), 3290–7.

Knowles, D., & Ghahramani, Z. (2010). NONPARAMETRIC BAYESIAN SPARSE FACTOR MODELS WITH APPLICATION TO GENE EXPRESSION MODELING. *The Annals of Applied Statistics*, **5**(2B), 1–21.

Koester-Eiserfunke, N., & Fischle, W. (2011). H3K9ME2/3 BINDING OF THE MBT DOMAIN PROTEIN LIN-61 IS ESSENTIAL FOR CAENORHABDITIS ELEGANS VULVA DEVELOPMENT. *PLoS Genetics*, **7**(3), e1002017.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X. S., & Ahringer, J. (2009).

DIFFERENTIAL CHROMATIN MARKING OF INTRONS AND EXPRESSED EXONS BY H3K36ME3. *Nature Genetics*, **41**(3), 376–381.

Koning, A. P. J. de, Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). REPETITIVE ELEMENTS MAY COMPRISE OVER TWO-THIRDS OF THE HUMAN GENOME. *PLoS Genetics*, **7**(12). doi:10.1371/JOURNAL.PGEN.1002384

Krause, M., & Hirsh, D. (1987). A TRANS-SPLICED LEADER SEQUENCE ON ACTIN MRNA IN C. ELEGANS. *Cell*, **49**, 753–761.

Kruesi, W. S., Core, L. J., Waters, C. T., Lis, J. T., & Meyer, B. J. (2013). CONDENSIN CONTROLS RECRUITMENT OF RNA POLYMERASE II TO ACHIEVE NEMATODE X-CHROMOSOME DOSAGE COMPENSATION. *ELife*, **2**, e00808.

Kumar, G. S., Chang, W., Xie, T., Patel, A., Zhang, Y., Wang, G. G., David, G., & Radhakrishnan, I. (2012). SEQUENCE REQUIREMENTS FOR COMBINATORIAL RECOGNITION OF HISTONE H3 BY THE MRG15 AND PF1 SUBUNITS OF THE RPD3S/SIN3S COREPRESSOR COMPLEX. *Journal of Molecular Biology*, **422**(4), 519–531.

Kuzmichev, A., Zhang, Y., Erdjument-Bromage, H., Tempst, P., & Reinberg, D. (2002). ROLE OF THE SIN3-HISTONE DEACETYLASE COMPLEX IN GROWTH REGULATION BY THE CANDIDATE TUMOR SUPPRESSOR P33ING1. *Molecular and Cellular Biology*, **22**(3), 835.

Kvon, E. Z., Stampfel, G., Yáñez-Cuna, J. O., Dickson, B. J., & Stark, A. (2012). HOT REGIONS FUNCTION AS PATTERNED DEVELOPMENTAL ENHANCERS AND HAVE A DISTINCT CIS-REGULATORY SIGNATURE. *Genes & Development*, **26**(9), 908–13.

Lange, U. C., Siebert, S., Wossidlo, M., Weiss, T., Ziegler-Birling, C., Walter, J., Torres-Padilla, M.-E., Daujat, S., & Schneider, R. (2013). DISSECTING THE ROLE OF H3K64ME3 IN MOUSE PERICENTROMERIC HETEROCHROMATIN. *Nature Communications*, **4**, 2233.

Latorre, I., Chesney, M. a, Garrigues, J. M., Stempor, P., Appert, A., Francesconi, M.,

Strome, S., & Ahringer, J. (2013). THE DREAM COMPLEX PROMOTES GENE BODY H2A . Z FOR TARGET REPRESSION. *Genes Dev*, **29**(5), 495–500.

Lee, H.-C., Gu, W., Shirayama, M., Youngman, E., Conte, D., & Mello, C. C. (2012). C. ELEGANS piRNAS MEDIATE THE GENOME-WIDE SURVEILLANCE OF GERMLINE TRANSCRIPTS. *Cell*, **150**(1), 78–87.

Lee, J.-H., & Skalnik, D. G. (2005). CPG-BINDING PROTEIN (CXXC FINGER PROTEIN 1) IS A COMPONENT OF THE MAMMALIAN SET1 HISTONE H3-LYS 4 METHYLTRANSFERASE COMPLEX, THE ANALOGUE OF THE YEAST SET1/COMPASS COMPLEX. *Journal of Biological Chemistry*, **280**(50), 41725–41731.

Lee, W. Y., Lee, D., Chung, W.-I., & Kwon, C. S. (2009). ARABIDOPSIS ING AND ALFIN1-LIKE PROTEIN FAMILIES LOCALIZE TO THE NUCLEUS AND BIND TO H3K4ME3/2 VIA PLANT HOMEODOMAIN FINGERS. doi:10.1111/j.1365-313X.2009.03795.x

Li, H., Ilin, S., Wang, W., Duncan, E. M., Wysocka, J., Allis, C. D., & Patel, D. J. (2006). MOLECULAR BASIS FOR SITE-SPECIFIC READ-OUT OF HISTONE H3K4ME3 BY THE BPTF PHD FINGER OF NURF. *Nature*, **442**(7098), 91–95.

Li, T., & Kelly, W. G. (2011). A ROLE FOR SET1/MLL-RELATED COMPONENTS IN EPIGENETIC REGULATION OF THE CAENORHABDITIS ELEGANS GERM LINE. *PLoS Genetics*, **7**(3), e1001349.

Liu, T., Ortiz, J. A., Taing, L., Meyer, C. A., Lee, B., Zhang, Y., Shin, H., Wong, S. S., Ma, J., Lei, Y., Pape, U. J., Poidinger, M., Chen, Y., Yeung, K., Brown, M., Turpaz, Y., & Liu, X. S. (2011a). CISTROME: AN INTEGRATIVE PLATFORM FOR TRANSCRIPTIONAL REGULATION STUDIES. *Genome Biology*, **12**(8), R83.

Liu, T., Rechtsteiner, A., Egelhofer, T. A., Vielle, A., Latorre, I., Cheung, M.-S. M.-S., Ercan, S., Ikegami, K., Jensen, M., Kolasinska-zwierz, P., Rosenbaum, H., Shin, H., Taing, S., Takasaki, T., Iniguez, A. L., Desai, A., Dernburg, A. F., Kimura, H., Lieb, J. D., Ahringer, J., Strome, S., & Liu, X. S. (2011b). BROAD CHROMOSOMAL

DOMAINS OF HISTONE MODIFICATION PATTERNS IN C. ELEGANS. *Genome Research*, **21**(2), 227–36.

Love, M. I., Huber, W., & Anders, S. (2014). MODERATED ESTIMATION OF FOLD CHANGE AND DISPERSION FOR RNA-SEQ DATA WITH DESEQ2. *Genome Biology*, **15**(12), 550.

Mao, H., Zhu, C., Zong, D., Weng, C., Yang, X., Huang, H., Liu, D., Feng, X., & Guang, S. (2015). THE NRDE PATHWAY MEDIATES SMALL-RNA-DIRECTED HISTONE H3 LYSINE 27 TRIMETHYLATION IN CAENORHABDITIS ELEGANS. *Current Biology*, **25**(18), 2398–2403.

Marin, I., Plata-Rengifo, P., Labrador, M., & Fontdevila, A. (1998). EVOLUTIONARY RELATIONSHIPS AMONG THE MEMBERS OF AN ANCIENT CLASS OF NON-LTR RETROTRANSPOSONS FOUND IN THE NEMATODE CAENORHABDITIS ELEGANS. *Molecular Biology and Evolution*, **15**(11), 1390–1402.

Martin, D. G. E., Baetz, K., Shi, X., Walter, K. L., MacDonald, V. E., Wlodarski, M. J., Gozani, O., Hieter, P., & Howe, L. (2006). THE YNG1P PLANT HOMEODOMAIN FINGER IS A METHYL-HISTONE BINDING MODULE THAT RECOGNIZES LYSINE 4-METHYLATED HISTONE H3. *Molecular and Cellular Biology*, **26**(21), 7871–9.

McMurchy, A. N., Stempor, P., Gaarenstroom, T., Wysolmerski, B., Dong, Y., Aussianikava, D., Appert, A., Huang, N., Kolasinska-Zwierz, P., Sapetschnig, A., Miska, E. A., & Ahringer, J. (2017). A TEAM OF HETEROCHROMATIN FACTORS COLLABORATES WITH SMALL RNA PATHWAYS TO COMBAT REPETITIVE ELEMENTS AND GERMLINE STRESS. *ELife*, **6**, e21666.

Megiorni, F., Cialfi, S., Dominici, C., Quattrucci, S., & Pizzuti, A. (2011). SYNERGISTIC POST-TRANSCRIPTIONAL REGULATION OF THE CYSTIC FIBROSIS TRANSMEMBRANE CONDUCTANCE REGULATOR (CFTR) BY MIR-101 AND MIR-494 SPECIFIC BINDING. *PloS One*, **6**(10), e26601.

Meister, P., Schott, S., Bedet, C., Xiao, Y., Rohner, S., Bodennec, S., Hudry, B., Molin,

L., Solari, F., Gasser, S. M., & Palladino, F. (2011). CAENORHABDITIS ELEGANS HETEROCHROMATIN PROTEIN 1 (HPL-2) LINKS DEVELOPMENTAL PLASTICITY, LONGEVITY AND LIPID METABOLISM. *Genome Biology*, **12**(12), R123.

Meléndez, A., & Greenwald, I. (2000). CAENORHABDITIS ELEGANS LIN-13, A MEMBER OF THE LIN-35 RB CLASS OF GENES INVOLVED IN VULVAL DEVELOPMENT, ENCODES A PROTEIN WITH ZINC FINGERS AND AN LXCXE MOTIF. *Genetics*, **155**(3), 1127–37.

Melters, D. P., Paliulis, L. V., Korf, I. F., & Chan, S. W. L. (2012). HOLOCENTRIC CHROMOSOMES: CONVERGENT EVOLUTION, MEIOTIC ADAPTATIONS, AND GENOMIC ANALYSIS. *Chromosome Research*, **20**(5), 579–593.

Naclerio, G., Cangiano, G., Coulson, A., Levitt, A., Ruvolo, V., & La Volpe, A. (1992). MOLECULAR AND GENOMIC ORGANIZATION OF CLUSTERS OF REPETITIVE DNA SEQUENCES IN CAENORHABDITIS ELEGANS. *Journal of Molecular Biology*, **226**(1), 159–68.

Nègre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., Kheradpour, P., Eaton, M. L., Loriaux, P., Sealfon, R., Li, Z., Ishii, H., Spokony, R. F., Chen, J., Hwang, L., Cheng, C., Auburn, R. P., Davis, M. B., Domanus, M., Shah, P. K., Morrison, C. A., Zieba, J., Suchy, S., Senderowicz, L., Victorsen, A., Bild, N. A., Grundstad, A. J., Hanley, D., MacAlpine, D. M., Mannervik, M., Venken, K., Bellen, H., White, R., Gerstein, M., Russell, S., Grossman, R. L., Ren, B., Posakony, J. W., Kellis, M., & White, K. P. (2011). A CIS-REGULATORY MAP OF THE DROSOPHILA GENOME. *Nature*, **471**(7339), 527–531.

Ni, J., Clark, K. J., Fahrenkrug, S. C., & Ekker, S. C. (2008). TRANSPOSON TOOLS HOPPING IN VERTEBRATES. *Briefings in Functional Genomics and Proteomics*, **7**(6), 444–453.

Ni, J. Z., Chen, E., & Gu, S. G. (2014). COMPLEX CODING OF ENDOGENOUS SIRNA, TRANSCRIPTIONAL SILENCING AND H3K9 METHYLATION ON NATIVE TARGETS OF GERMLINE NUCLEAR RNAI IN C. ELEGANS. *BMC Genomics*, **15**(6669), 1157.

Niu, W., Hart, G. T., & Marcotte, E. M. (2011). HIGH-THROUGHPUT IMMUNOFLUORESCENCE MICROSCOPY USING YEAST SPHEROPLAST CELL-BASED MICROARRAYS. *Methods in Molecular Biology (Clifton, N.J.)*, **706**, 83–95.

Ooi, L., & Wood, I. C. (2007). CHROMATIN CROSSTALK IN DEVELOPMENT AND DISEASE: LESSONS FROM REST. *Nature Reviews. Genetics*, **8**(July), 544–554.

Oosumi, T., Garlick, B., & Belknap, W. R. (1996). IDENTIFICATION OF PUTATIVE NONAUTONOMOUS TRANSPOSABLE ELEMENTS ASSOCIATED WITH SEVERAL TRANSPOSON FAMILIES IN CAENORHABDITIS ELEGANS. *Journal of Molecular Evolution*, **43**(1), 11–8.

Passannante, M., Marti, C.-O., Pfefferli, C., Moroni, P. S., Kaeser-Pebernard, S., Puoti, A., Hunziker, P., Wicky, C., & Müller, F. (2010a). DIFFERENT MI-2 COMPLEXES FOR VARIOUS DEVELOPMENTAL FUNCTIONS IN CAENORHABDITIS ELEGANS. *PLoS ONE*, **5**(10), e13681.

Passannante, M., Marti, C.-O., Pfefferli, C., Moroni, P. S., Kaeser-Pebernard, S., Puoti, A., Hunziker, P., Wicky, C., & Müller, F. (2010b). DIFFERENT MI-2 COMPLEXES FOR VARIOUS DEVELOPMENTAL FUNCTIONS IN CAENORHABDITIS ELEGANS. *PLoS ONE*, **5**(10), e13681.

Pinskaya, M., Gourvennec, S., & Morillon, A. (2009). H3 LYSINE 4 DI- AND TRI-METHYLATION DEPOSITED BY CRYPTIC TRANSCRIPTION ATTENUATES PROMOTER ACTIVATION. *The EMBO Journal*, **28**(12), 1697–1707.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). DEEPTOOLS: A FLEXIBLE PLATFORM FOR EXPLORING DEEP-SEQUENCING DATA. *Nucleic Acids Research*, **42**(W1). doi:10.1093/nar/gku365

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). DETECTING NOVEL ASSOCIATIONS IN LARGE DATA SETS. *Science (New York, N.Y.)*, **334**(6062), 1518–24.

Riddle, D. L., Blumenthal, T., Meyer, B. J., & Priess, J. R. (1997). C. ELEGANS II, Cold Spring Harbor Laboratory Press. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK19997/

Robertson, A. G., Bilenky, M., Tam, A., Zhao, Y., Zeng, T., Thiessen, N., Cezard, T., Fejes, A. P., Wederell, E. D., Cullum, R., Euskirchen, G., Krzywinski, M., Birol, I., Snyder, M., Hoodless, P. A., Hirst, M., Marra, M. A., & Jones, S. J. M. (2008). GENOME-WIDE RELATIONSHIP BETWEEN HISTONE H3 LYSINE 4 MONO- AND TRI-METHYLATION AND TRANSCRIPTION FACTOR BINDING. *Genome Research*, **18**(12), 1906–17.

Rosenzweig, B., Liao, L. W., & Hirsh, D. (1983). SEQUENCE OF THE C. ELEGANS TRANSPOSABLE ELEMENT TC1. *Nucleic Acids Research*, **11**(12), 4201–9.

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., Brown, G. D., Gojis, O., Ellis, I. O., Green, A. R., Ali, S., Chin, S.-F., Palmieri, C., Caldas, C., & Carroll, J. S. (2012). DIFFERENTIAL OESTROGEN RECEPTOR BINDING IS ASSOCIATED WITH CLINICAL OUTCOME IN BREAST CANCER. *Nature*, **481**(7381), 389–393.

Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., Lin, M. F., Washietl, S., Arshinoff, B. I., Ay, F., Meyer, P. E., Robine, N., Washington, N. L., Di Stefano, L., Berezikov, E., Brown, C. D., Candeias, R., Carlson, J. W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M. Y., Will, S., Alekseyenko, A. A., Artieri, C., Booth, B. W., Brooks, A. N., Dai, Q., Davis, C. A., Duff, M. O., Feng, X., Gorchakov, A. A., Gu, T., Henikoff, J. G., Kapranov, P., Li, R., MacAlpine, H. K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S. K., Riddle, N. C., Sakai, A., Samsonova, A., Sandler, J. E., Schwartz, Y. B., Sher, N., Spokony, R., Sturgill, D., van Baren, M., Wan, K. H., Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S. E., Brent,

M. R., Cherbas, L., Elgin, S. C. R., Gingeras, T. R., Grossman, R., Hoskins, R. A., Kaufman, T. C., Kent, W., Kuroda, M. I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J. W., Ren, B., Russell, S., Cherbas, P., Graveley, B. R., Lewis, S., Micklem, G., Oliver, B., Park, P. J., Celniker, S. E., Henikoff, S., Karpen, G. H., Lai, E. C., MacAlpine, D. M., Stein, L. D., White, K. P., Kellis, M., Acevedo, D., Auburn, R., Barber, G., Bellen, H. J., Bishop, E. P., Bryson, T. D., Chateigner, A., Chen, J., Clawson, H., Comstock, C. L. G., Contrino, S., DeNapoli, L. C., Ding, Q., Dobin, A., Domanus, M. H., Drenkow, J., Dudoit, S., Dumais, J., Eng, T., Fagegaltier, D., Gadel, S. E., Ghosh, S., Guillier, F., Hanley, D., Hannon, G. J., Hansen, K. D., Heinz, E., Hinrichs, A. S., Hirst, M., Jha, S., Jiang, L., Jung, Y. L., Kashevsky, H., Kennedy, C. D., Kephart, E. T., Langton, L., Lee, O.-K., Li, S., Li, Z., Lin, W., Linder-Basso, D., Lloyd, P., Lyne, R., Marchetti, S. E., Marra, M., Mattiuzzo, N. R., McKay, S., Meyer, F., Miller, D., Miller, S. W., Moore, R. A., Morrison, C. A., Prinz, J. A., Rooks, M., Moore, R., Rutherford, K. M., Ruzanov, P., Scheftner, D. A., Senderowicz, L., Shah, P. K., Shanower, G., Smith, R., Stinson, E. O., Suchy, S., Tenney, A. E., Tian, F., Venken, K. J. T., Wang, H., White, R., Wilkening, J., Willingham, A. T., Zaleski, C., Zha, Z., Zhang, D., Zhao, Y., & Zieba, J. (2010). IDENTIFICATION OF FUNCTIONAL ELEMENTS AND REGULATORY CIRCUITS BY DROSOPHILA MODENCODE. *Science*, **330**(6012), 1787–1797.

Saksouk, N., Simboeck, E., & Déjardin, J. (2015). CONSTITUTIVE HETEROCHROMATIN FORMATION AND TRANSCRIPTION IN MAMMALS. *Epigenetics & Chromatin*, **8**, 3.

Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., Emre, N. C. T., Schreiber, S. L., Mellor, J., & Kouzarides, T. (2002). ACTIVE GENES ARE TRI-METHYLATED AT K4 OF HISTONE H3. *Nature*, **419**(6905), 407–411.

Santos-Rosa, H., Schneider, R., Bernstein, B. E., Karabetsou, N., Morillon, A., Weise, C., Schreiber, S. L., Mellor, J., & Kouzarides, T. (2003). METHYLATION OF

HISTONE H3 K4 MEDIATES ASSOCIATION OF THE ISW1P ATPASE WITH CHROMATIN. *Molecular Cell*, **12**(5), 1325–1332.

Sarachana, T., Zhou, R., Chen, G., Manji, H. K., & Hu, V. W. (2010). INVESTIGATION OF POST-TRANSCRIPTIONAL GENE REGULATORY NETWORKS ASSOCIATED WITH AUTISM SPECTRUM DISORDERS BY MICRORNA EXPRESSION PROFILING OF LYMPHOBLASTOID CELL LINES. *Genome Medicine*, **2**(4), 23.

Scheifele, L. Z., Cost, G. J., Zupancic, M. L., Caputo, E. M., & Boeke, J. D. (n.d.). RETROTRANSPOSON OVERDOSE AND GENOME INTEGRITY. Retrieved from http://www.pnas.org/content/pnas/106/33/13927.full.pdf

Schott, S., Coustham, V., Simonet, T., Bedet, C., & Palladino, F. (2006). UNIQUE AND REDUNDANT FUNCTIONS OF C. ELEGANS HP1 PROTEINS IN POST-EMBRYONIC DEVELOPMENT. *Developmental Biology*, **298**(1), 176–187.

Sha, K., Gu, S. G., Pantalena-Filho, L. C., Goh, A., Fleenor, J., Blanchard, D., Krishna, C., & Fire, A. (2010). DISTRIBUTED PROBING OF CHROMATIN STRUCTURE IN VIVO REVEALS PERVASIVE CHROMATIN ACCESSIBILITY FOR EXPRESSED AND NON-EXPRESSED GENES DURING TISSUE DIFFERENTIATION IN C. ELEGANS. *BMC Genomics*, **11**(1), 465.

Shen, L., Shao, N., Liu, X., & Nestler, E. (2014). NGS.PLOT: QUICK MINING AND VISUALIZATION OF NEXT-GENERATION SEQUENCING DATA BY INTEGRATING GENOMIC DATABASES. *BMC Genomics*, **15**(1), 284.

Shi, X., Hong, T., Walter, K. L., Ewalt, M., Michishita, E., Hung, T., Carney, D., Peña, P., Lan, F., Kaadige, M. R., Lacoste, N., Cayrou, C., Davrazou, F., Saha, A., Cairns, B. R., Ayer, D. E., Kutateladze, T. G., Shi, Y., Côté, J., Chua, K. F., & Gozani, O. (2006). ING2 PHD DOMAIN LINKS HISTONE H3 LYSINE 4 METHYLATION TO ACTIVE GENE REPRESSION. *Nature*, **442**(7098), 96–99.

Simonet, T., Dulermo, R., Schott, S., & Palladino, F. (2007). ANTAGONISTIC FUNCTIONS OF SET-2/SET1 AND HPL/HP1 PROTEINS IN C. ELEGANS DEVELOPMENT.

Developmental Biology, **312**(1), 367–383.

Simpson, V. J., Johnson, T. E., & Hammen, R. F. (1986). CAENORHABDITIS ELEGANS DNA DOES NOT CONTAIN 5-METHYLCYTOSINE AT ANY TIME DURING DEVELOPMENT OR AGING. *Nucleic Acids Research*, **14**(16), 6711–6719.

Sofueva, S., & Hadjur, S. (2012). COHESIN-MEDIATED CHROMATIN INTERACTIONS--INTO THE THIRD DIMENSION OF GENE REGULATION. *Briefings in Functional Genomics*, **11**(3), 205–16.

Spencer, W. C., Zeller, G., Watson, J. D., Henz, S. R., Watkins, K. L., McWhirter, R. D., Petersen, S., Sreedharan, V. T., Widmer, C., Jo, J., Reinke, V., Petrella, L., Strome, S., Von Stetina, S. E., Katz, M., Shaham, S., Rätsch, G., & Miller, D. M. (2011). A SPATIAL AND TEMPORAL MAP OF C. ELEGANS GENE EXPRESSION. *Genome Research*, **21**(2), 325–41.

Spieth, J., Brooke, G., Kuersten, S., Lea, K., & Blumenthal, T. (1993). OPERONS IN C. ELEGANS: POLYCISTRONIC MRNA PRECURSORS ARE PROCESSED BY TRANS-SPLICING OF SL2 TO DOWNSTREAM CODING REGIONS. *Cell*, **73**, 521–532.

Spieth, J., Lawson, D., Davis, P., Williams, G., & Howe, K. (2013). OVERVIEW OF GENE STRUCTURE IN C. ELEGANS. *WormBook*, 1–30.

Stempor, P. (2014, August 26). RBEADS - THE R IMPLEMENTATION OF BIAS ELIMINATION ALGORITHM FOR DEEP SEQUENCING. doi:10.5281/zenodo.11427

Stempor, P., & Ahringer, J. (2016). SEQPLOTS - INTERACTIVE SOFTWARE FOR EXPLORATORY DATA ANALYSES, PATTERN DISCOVERY AND VISUALIZATION IN GENOMICS. *Wellcome Open Research*, **1**, 14.

Sulli, G., Di Micco, R., & di Fagagna, F. d'Adda. (2012). CROSSTALK BETWEEN CHROMATIN STATE AND DNA DAMAGE RESPONSE IN CELLULAR SENESCENCE AND CANCER. *Nature Reviews Cancer*, **12**(10), 709–720.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). MEASURING AND TESTING DEPENDENCE BY CORRELATION OF DISTANCES. *The Annals of Statistics*, **35**(6),

2769–2794.

Tan, B. G., Vijgenboom, E., & Worrall, J. A. R. (2014). CONFORMATIONAL AND THERMODYNAMIC HALLMARKS OF DNA OPERATOR SITE SPECIFICITY IN THE COPPER SENSITIVE OPERON REPRESSOR FROM STREPTOMYCES LIVIDANS. *Nucleic Acids Research*, **42**(2), 1326–40.

Tan, J. H., & Fraser, A. G. (2017). THE COMBINATORIAL CONTROL OF ALTERNATIVE SPLICING IN C. ELEGANS. *PLOS Genetics*, **13**(11), e1007033.

Tate, C. M., Lee, J. H., & Skalnik, D. G. (2010). CXXC FINGER PROTEIN 1 RESTRICTS THE SETD1A HISTONE H3K4 METHYLTRANSFERASE COMPLEX TO EUCHROMATIN. *FEBS Journal*, **277**(1), 210–223.

Taverna, S. D., Ilin, S., Rogers, R. S., Tanny, J. C., Lavender, H., Li, H., Baker, L., Boyle, J., Blair, L. P., Chait, B. T., Patel, D. J., Aitchison, J. D., Tackett, A. J., & Allis, C. D. (2006). YNG1 PHD FINGER BINDING TO H3 TRIMETHYLATED AT K4 PROMOTES NUA3 HAT ACTIVITY AT K14 OF H3 AND TRANSCRIPTION AT A SUBSET OF TARGETED ORFS. *Molecular Cell*, **24**(5), 785–796.

Teif, V. B., & Rippe, K. (2009). PREDICTING NUCLEOSOME POSITIONS ON THE DNA: COMBINING INTRINSIC SEQUENCE PREFERENCES AND REMODELER ACTIVITIES. *Nucleic Acids Research*, **37**(17), 5641–55.

Thomas, J. H., Ceol, C. J., Schwartz, H. T., & Horvitz, H. R. (2003). NEW GENES THAT INTERACT WITH LIN-35 RB TO NEGATIVELY REGULATE THE LET-60 RAS PATHWAY IN CAENORHABDITIS ELEGANS. *Genetics*, **164**(1), 135–51.

Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A. R. W., Deaton, A., Andrews, R., James, K. D., Turner, D. J., Illingworth, R., & Bird, A. (2010). CPG ISLANDS INFLUENCE CHROMATIN STRUCTURE VIA THE CPG-BINDING PROTEIN CFP1. *Nature*, **464**(7291), 1082–1086.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S.,

Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. a, Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., & Stamatoyannopoulos, J. a. (2012). THE ACCESSIBLE CHROMATIN LANDSCAPE OF THE HUMAN GENOME. *Nature*, **489**(7414), 75–82.

Towbin, B. D., González-Aguilera, C., Sack, R., Gaidatzis, D., Kalck, V., Meister, P., Askjaer, P., & Gasser, S. M. (2012). STEP-WISE METHYLATION OF HISTONE H3K9 POSITIONS HETEROCHROMATIN AT THE NUCLEAR PERIPHERY. *Cell*, **150**(5), 934–947.

Tsurumi, A., & Li, W. X. (2012). GLOBAL HETEROCHROMATIN LOSS: A UNIFYING THEORY OF AGING? *Epigenetics*, **7**(7), 680–8.

Valk, T. van der, Vezzi, F., Ormestad, M., Dalen, L., & Guschanski, K. (2018). ESTIMATING THE RATE OF INDEX HOPPING ON THE ILLUMINA HISEQ X PLATFORM. *BioRxiv*, 179028.

van Dongen, S., Abreu-Goodger, C., & Enright, A. J. (2008). DETECTING MICRORNA BINDING AND SIRNA OFF-TARGET EFFECTS FROM EXPRESSION DATA. *Nature Methods*, **5**(12), 1023–5.

Vavouri, T., & Lehner, B. (2012). HUMAN GENES WITH CPG ISLAND PROMOTERS HAVE A DISTINCT TRANSCRIPTION-ASSOCIATED CHROMATIN ORGANIZATION. *Genome Biology*, **13**(11), R110.

Vermeulen, M., Eberl, H. C., Matarese, F., Marks, H., Denissov, S., Butter, F., Lee, K. K., Olsen, J. V., Hyman, A. A., Stunnenberg, H. G., & Mann, M. (2010).

QUANTITATIVE INTERACTION PROTEOMICS AND GENOME-WIDE PROFILING OF EPIGENETIC HISTONE MARKS AND THEIR READERS. *Cell*, **142**(6), 967–980.

Villeponteau, B. (1997). THE HETEROCHROMATIN LOSS MODEL OF AGING. *Experimental Gerontology*, **32**(4–5), 383–94.

von Zelewsky, T., Palladino, F., Brunschwig, K., Tobler, H., Hajnal, A., & Müller, F. (2000). THE C. ELEGANS MI-2 CHROMATIN-REMODELLING PROTEINS FUNCTION IN VULVAL CELL FATE DETERMINATION. *Development (Cambridge, England)*, **127**(24), 5277–84.

Voo, K. S., Carlone, D. L., Jacobsen, B. M., Flodin, A., & Skalnik, D. G. (2000). CLONING OF A MAMMALIAN TRANSCRIPTIONAL ACTIVATOR THAT BINDS UNMETHYLATED CPG MOTIFS AND SHARES A CXXC DOMAIN WITH DNA METHYLTRANSFERASE, HUMAN TRITHORAX, AND METHYL-CPG BINDING DOMAIN PROTEIN 1. *Molecular and Cellular Biology*, **20**(6), 2108–21.

Wang, P., Lin, C., Smith, E. R., Guo, H., Sanderson, B. W., Wu, M., Gogol, M., Alexander, T., Seidel, C., Wiedemann, L. M., Ge, K., Krumlauf, R., & Shilatifard, A. (2009). GLOBAL ANALYSIS OF H3K4 METHYLATION DEFINES MLL FAMILY MEMBER TARGETS AND POINTS TO A ROLE FOR MLL1-MEDIATED H3K4 METHYLATION IN THE REGULATION OF TRANSCRIPTIONAL INITIATION BY RNA POLYMERASE II. *Molecular and Cellular Biology*, **29**(22), 6074–6085.

Ward, M. C., Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., Stark, R., Pan, Q., Schwalie, P. C., Menon, S., Lukk, M., Watt, S., Thybert, D., Kutter, C., Kirschner, K., Flicek, P., Blencowe, B. J., & Odom, D. T. (2013). LATENT REGULATORY POTENTIAL OF HUMAN-SPECIFIC REPETITIVE ELEMENTS. *Molecular Cell*, **49**(2), 262–72.

Weick, E.-M. M., Sarkies, P., Silva, N., Chen, R. a., Moss, S. M. M., Cording, A. C., Ahringer, J., Martinez-Perez, E., & Miska, E. a. (2014). PRDE-1 IS A NUCLEAR FACTOR ESSENTIAL FOR THE BIOGENESIS OF RUBY MOTIF-DEPENDENT PIRNAS IN C.

ELEGANS. *Genes & Development*, **28**(7), 783–96.

Williams, T., Kelley, C., & many others. (2013, April). GNUPLOT 4.6: AN INTERACTIVE PLOTTING PROGRAM. Retrieved from http://www.gnuplot.info/

Williamson, I., Berlivet, S., Eskeland, R., Boyle, S., Illingworth, R. S., Paquette, D., Dostie, J., & Bickmore, W. A. (2014). SPATIAL GENOME ORGANIZATION: CONTRASTING VIEWS FROM CHROMOSOME CONFORMATION CAPTURE AND FLUORESCENCE IN SITU HYBRIDIZATION. *Genes & Development*, **28**(24), 2778–2791.

Wu, X., Shi, Z., Cui, M., Han, M., & Ruvkun, G. (2012). REPRESSION OF GERMLINE RNAI PATHWAYS IN SOMATIC CELLS BY RETINOBLASTOMA PATHWAY CHROMATIN COMPLEXES. *PLoS Genetics*, **8**(3). doi:10.1371/JOURNAL.PGEN.1002542

Wysocka, J., Myers, M. P., Laherty, C. D., Eisenman, R. N., & Herr, W. (2003). HUMAN SIN3 DEACETYLASE AND TRITHORAX-RELATED SET1/ASH2 HISTONE H3-K4 METHYLTRANSFERASE ARE TETHERED TOGETHER SELECTIVELY BY THE CELL-PROLIFERATION FACTOR HCF-1. *Genes and Development*, **17**(7), 896–911.

Wysocka, J., Swigut, T., Xiao, H., Milne, T. A., Kwon, S. Y., Landry, J., Kauer, M., Tackett, A. J., Chait, B. T., Badenhorst, P., Wu, C., & Allis, C. D. (2006). A PHD FINGER OF NURF COUPLES HISTONE H3 LYSINE 4 TRIMETHYLATION WITH CHROMATIN REMODELLING. *Nature*, **442**(7098), 86–90.

Xiao, Y., Bedet, C., Robert, V. J. P., Simonet, T., Dunkelbarger, S., Rakotomalala, C., Soete, G., Korswagen, H. C., Strome, S., & Palladino, F. (2011). CAENORHABDITIS ELEGANS CHROMATIN-ASSOCIATED PROTEINS SET-2 AND ASH-2 ARE DIFFERENTIALLY REQUIRED FOR HISTONE H3 LYS 4 METHYLATION IN EMBRYOS AND ADULT GERM CELLS. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(20), 8305–10.

Xu, C., Bian, C., Lam, R., Dong, A., & Min, J. (2011). THE STRUCTURAL BASIS FOR SELECTIVE BINDING OF NON-METHYLATED CPG ISLANDS BY THE CFP1 CXXC

DOMAIN. *Nature Communications*, **2**(1), 227.

Yanez-Cuna, J. O., Arnold, C. D., Stampfel, G., Boryn, L. M., Gerlach, D., Rath, M., & Stark, A. (2014). DISSECTION OF THOUSANDS OF CELL TYPE-SPECIFIC ENHANCERS IDENTIFIES DINUCLEOTIDE REPEAT MOTIFS AS GENERAL ENHANCER FEATURES. *Genome Research*. doi:10.1101/gr.169243.113

Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., & Gerstein, M. (2012). CLASSIFICATION OF HUMAN GENOMIC REGIONS BASED ON EXPERIMENTALLY DETERMINED BINDING SITES OF MORE THAN 100 TRANSCRIPTION-RELATED FACTORS. *Genome Biology*, **13**(9), R48.

Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A., & Majewski, I. J. (2011). CHIP-SEQ ANALYSIS REVEALS DISTINCT H3K27ME3 PROFILES THAT CORRELATE WITH TRANSCRIPTIONAL ACTIVITY. *Nucleic Acids Research*, **39**(17), 7415–27.

Yücel, D., Hoe, M., Llamosas, E., Kant, S., Jamieson, C., Young, P. A., Crossley, M., & Nicholas, H. R. (2014). SUMV-1 ANTAGONIZES THE ACTIVITY OF SYNTHETIC MULTIVULVA GENES IN CAENORHABDITIS ELEGANS. *Developmental Biology*, **392**(2), 266–282.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). MODEL-BASED ANALYSIS OF CHIP-SEQ (MACS). *Genome Biology*, **9**(9), R137.

8 APPENDICES

8.1 List of relevant publications I co-authored

- Beurton, F., **Stempor, P.**, Bedet, C., Caron, M., Appert, A., Herbette, M., Polveche, H., Couté, Y., Spichty, M., Chen, R., Huang, N., Dong, Y., Cluet, D., Ahringer, J., & Palladino, F. (2018). **PHYSICAL AND FUNCTIONAL INTERACTION BETWEEN THE SET1/COMPASS COMPLEX COMPONENT CFP-1/CXXC AND A SIN3S HDAC COMPLEX.** *In Pess*.
- Janes, J., Dong, Y., Schoof, M., Serizay, J., Appert, A., Cerrato, C., Woodbury, C., Chen, R., Gemma, C., Huang, N., Kissiov, D., **Stempor, P.**, Steward, A., Zeiser, E., Sauer, S., & Ahringer, J. (2018). **CHROMATIN ACCESSIBILITY DYNAMICS ACROSS C. ELEGANS DEVELOPMENT AND AGEING.** *BioRxiv*, 279158.
- McMurchy, A. N., **Stempor, P.**, Gaarenstroom, T., Wysolmerski, B., Dong, Y., Aussianikava, D., Appert, A., Huang, N., Kolasinska-Zwierz, P., Sapetschnig, A., Miska, E. A., & Ahringer, J. (2017). **A TEAM OF HETEROCHROMATIN FACTORS COLLABORATES WITH SMALL RNA PATHWAYS TO COMBAT REPETITIVE ELEMENTS AND GERMLINE STRESS.** *ELife*, **6**, e21666.
- **Stempor, P.**, & Ahringer, J. (2016). SEQPLOTS - **INTERACTIVE SOFTWARE FOR EXPLORATORY DATA ANALYSES, PATTERN DISCOVERY AND VISUALIZATION IN GENOMICS.** *Wellcome Open Research*, **1**, 14.

- Evans, K. J., Huang, N., **Stempor, P.**, Chesney, M. A., Down, T. A., & Ahringer, J. (2016). **STABLE CAENORHABDITIS ELEGANS CHROMATIN DOMAINS SEPARATE BROADLY EXPRESSED AND DEVELOPMENTALLY REGULATED GENES.** *Proceedings of the National Academy of Sciences*, **113**(45), E7020–E7029.
- Chen, R. A.-J., **Stempor, P.**, Down, T. A., Zeiser, E., Feuer, S. K., & Ahringer, J. (2014). **EXTREME HOT REGIONS ARE CPG-DENSE PROMOTERS IN C. ELEGANS AND HUMANS.** *Genome Research*, **24**(7), 1138–46.
- **Stempor, P.** (2014, August 26). RBEADS - THE R IMPLEMENTATION OF BIAS ELIMINATION ALGORITHM FOR DEEP SEQUENCING. doi:10.5281/zenodo.11427
- Ho, J. W. K., Jung, Y. L., Liu, T., Alver, B. H., Lee, S., Ikegami, K., Sohn, K., Minoda, A., Tolstorukov, M. Y., Appert, A., Parker, S. C. J., Gu, T., Kundaje, A., Riddle, N. C., Bishop, E., Egelhofer, T. a, Hu, S., Alekseyenko, A. a, Rechtsteiner, A., Asker, D., Belsky, J. a, Bowman, S. K., Chen, Q. B., Chen, R. a, Day, D. S., Dong, Y., Dose, A. C., Duan, X., Epstein, C. B., Ercan, S., Feingold, E. a, Ferrari, F., Garrigues, J. M., Gehlenborg, N., Good, P. J., Haseley, P., He, D., Herrmann, M., Hoffman, M. M., Jeffers, T. E., Kharchenko, P. V, Kolasinska-zwierz, P., Kotwaliwale, C. V, Kumar, N., Langley, S. a, Larschan, E. N., Latorre, I., Libbrecht, M. W., Lin, X., Park, R., Pazin, M. J., Pham, H. N., Plachetka, A., Qin, B., Schwartz, Y. B., Shoresh, N., **Stempor, P.**, Vielle, A., Wang, C., Whittle, C. M., Xue, H., Kingston, R. E., Kim, J. H., Bernstein, B. E., Dernburg, A. F., Pirrotta, V., Kuroda, M. I., Noble, W. S., Tullius, T. D., Kellis, M., Macalpine, D. M., Strome, S., Elgin, S. C. R., & Ad, N. (2014). **COMPARATIVE ANALYSIS OFMETAZOAN CHROMATIN ORGANIZATION.** *Nature*, **512**(7515), 449–452.

- Chen, R. A.-J. J., Down, T. a., **Stempor, P**., Chen, Q. B., Egelhofer, T. a., Hillier, L. W., Jeffers, T. E., & Ahringer, J. (2013). **THE LANDSCAPE OF RNA POLYMERASE II TRANSCRIPTION INITIATION IN C. ELEGANS REVEALS PROMOTER AND ENHANCER ARCHITECTURES.** *Genome Research*, **23**(8), 1339–47.
- Latorre, I., Chesney, M. a, Garrigues, J. M., **Stempor, P.**, Appert, A., Francesconi, M., Strome, S., & Ahringer, J. (2013). **THE DREAM COMPLEX PROMOTES GENE BODY H2A . Z FOR TARGET REPRESSION.** *Genes Dev*, **29**(5), 495–500.

8.2 Additional ChIP-seq profile analyses for HDA-1 in *set-2* and *sin-3, and* SIN-3 in *set-2*

Summarizing results from previous chapters I have found that in HiConf CFP-1 binding sites:

- HDA-1 in *cfp-1* is significantly reduced
- SIN-3 in *cfp-1* is significantly reduced

Here I present further analyses, that yielded inconclusive results due to possible technical problems. HDA-1 in *set-2* show no conclusive change due to replicates mismatch. In summary:

- HDA-1 in *set-2* show no conclusive change due to replicates mismatch
- HDA-1 in *sin-3* is increased on average, but non-significant
- SIN-3 at *set-2* is reduced, but overlap between mutant replicates is weak

We will do additional experiments to confirm or reject these findings.

8.2.1 SIN-3 abundance seems reduced on high confidence CFP-1 regions in *set-2* mutant strains

Next, I continued analysing SIN-3 signal on signal on CFP-1 binding sites, this time in *set-2* background. I see reduction in HiConf (C1) cluster and no reduction in LoConf peaks (C2, **Figure 121**). However, replicates aa163 and aa164 show quite distinct difference in signal straight both on HiConf and LoConf. Nevertheless, both mutant replicates show weaker enrichment in cluster C1 in comparison to wild type. Both observations were confirmed by signal quantifications (**Figure 121**). Finally, I have tested that the loss of SIN-3 is significant in any replicates combination (**Figure 122**A) and in replicates combined (**Figure 122**B). On average, C2 shows no significant gain in CFP-1 signal. I conclude that in bulk analyses SIN-3 is reduced in HiConf CFP-1 sites in *set-2* mutant background. However, the mismatch between mutant replicates poised

me to perform differential binding analyses, in order to assess the consistency of this effect.

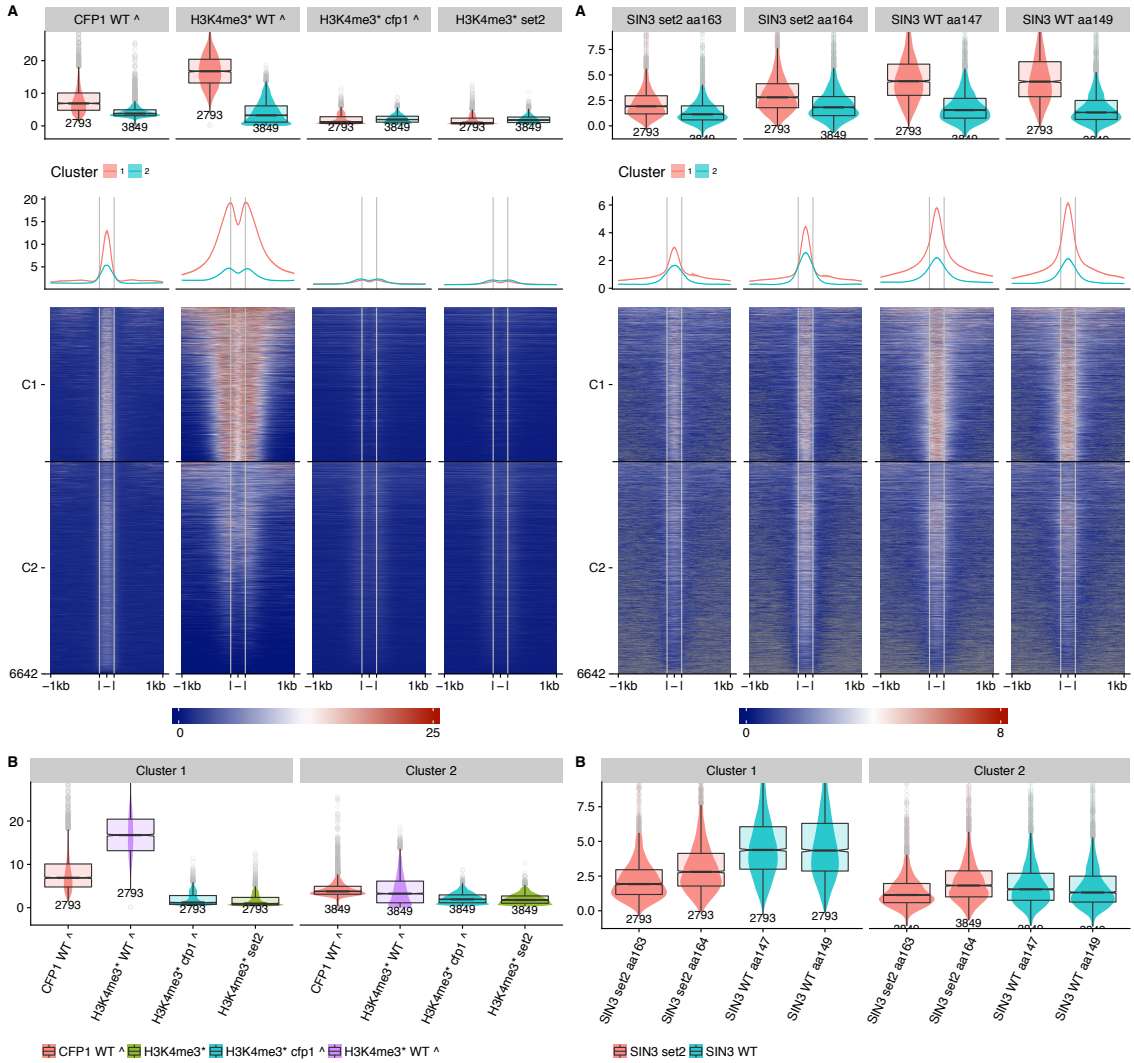


Figure 121 Quantification of SIN-3 in ChIP-seq in WT and *set-2* background shown in CFP-1 HiConf (C1) and LoConf (C2) peaks in context of CFP-1 signal and H3K3me3 in WT, cfp-1 and set-2 backgrounds. Panel A-left shows boxplot quantifications, profile plot quantifications and heatmaps for CFP-1 and H3K4me3 in WT, *cfp-1* and *set-2* mutant backgrounds. Panel A-right shows same plots for SIN-3 in WT and *set-2* strains. Panels B-left and B-right show same quantifications as boxplots on panel A, but are arranged in the way that makes it easy to compare replicates rather than clusters.

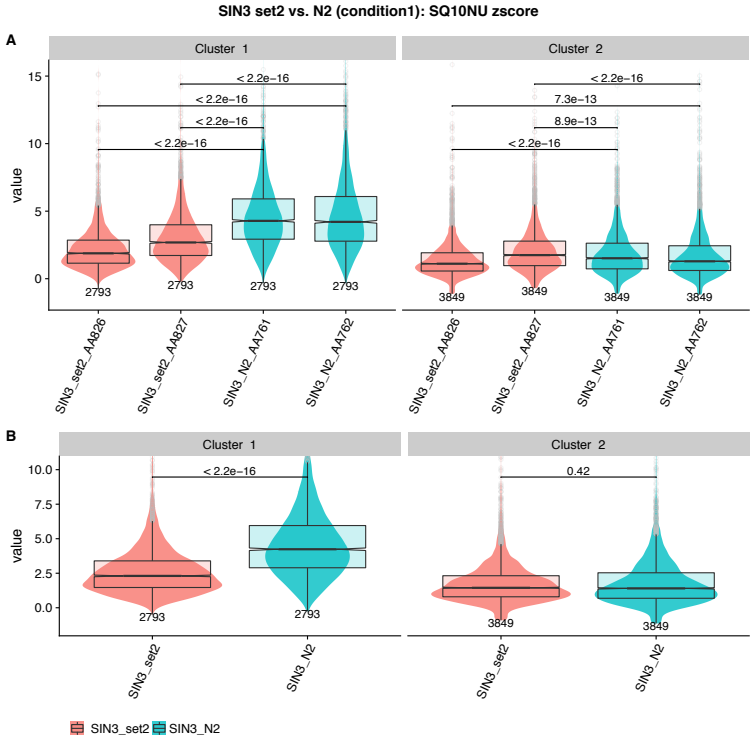


Figure 122 Significance testing for difference between SIN-3 in WT and *cfp-1* backgrounds. P-values are estimated using Mann-Whitney U test – a nonparametric method, which tests whether there is a location shift between two distributions (Bauer 1972).

8.2.2 HDA-1 signal change on CFP-1 bound regions in *set-2* and *sin-3* mutant strains is inconclusive

Finally, I have assessed HDA-1 signal change between set-2 and sin-3 mutants using the same regions. In both cases it is difficult to draw final conclusions because replicates show poor match (*sin-3*) or complete mismatch (*set-2*). In *set-2* experiment signal HDA-1 one replicate is lower, while in second is higher than in both WT replicates (**Figure 123**). Since, in each experiment we are sampling from big populations of *C. elegans*, I interpret this as technical issue with ChIP experiment rather than biological effect.

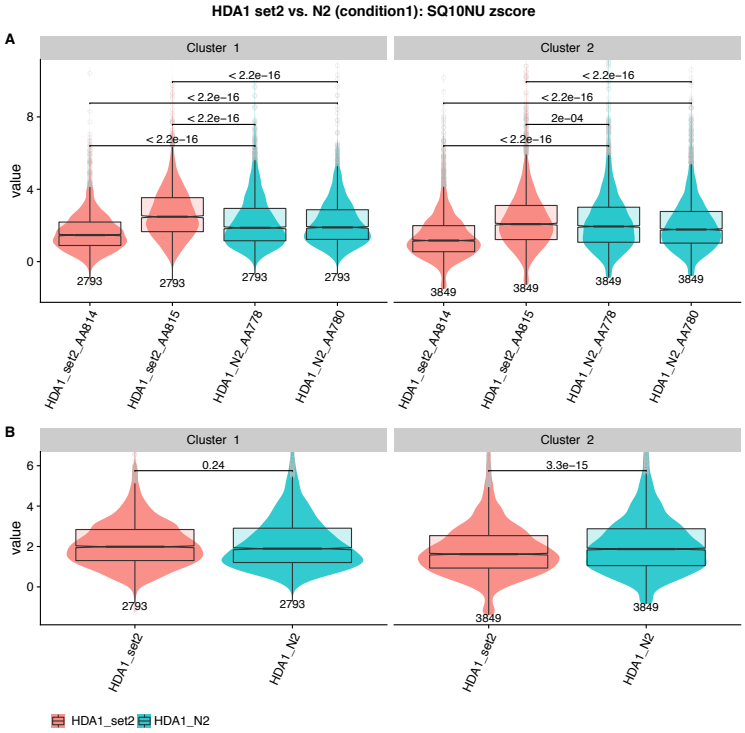


Figure 123 Significance testing for difference between HDA-1 in WT and *set-2* backgrounds. P-values are estimated using Mann-Whitney U test – a nonparametric method, which tests whether there is a location shift between two distributions (Bauer 1972).

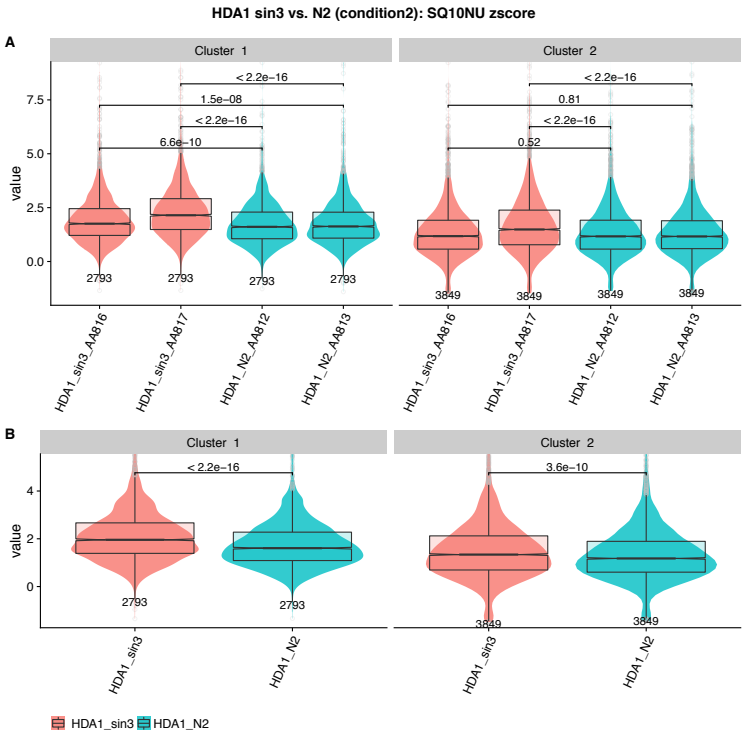


Figure 124 Significance testing for difference between HDA-1 in WT and *sin-3* backgrounds. P-values are estimated using Mann-Whitney U test – a nonparametric method, which test whether there is a location shift between two distributions (Bauer 1972).

In *sin-3* background replicates are better matched - both signals are higher in *sin-3* mutant in C1. Also, one replicate of *sin-3* experiments shows higher signal in C2, and second one is at similar level as WT samples. This would suggest a gain of HDA-1 signal in both clusters. However, in both clusters the signal of AA816 *sin-3* replicate is nearly as low as in WT. Considering this, above observation should be treated as possible outcome of technical, rather than biological factors and should be verified with further replicates.

8.3 Assessing cross-reactivity antibodies H3K9me antibodies

Following western blots are testing for cross-reactivity in H3K9me2 and H3K9me9 antibodies. WA 309-34839 H3K9me2 is a monoclonal antibody developed by Hiroshi Kimura (Kimura *et al.* 2008) and provided by Wako. AB8898 H3K9me2 antibody is a polyclonal antibody (animal lot number 339901) provided by Abcam. Western blasts show no cross-reactivity for H3K9me2 antibody (**Figure 125**), and vey week cross-reactivity with H3K9me2 peptide for H3K9me3 antibody (**Figure 126**).

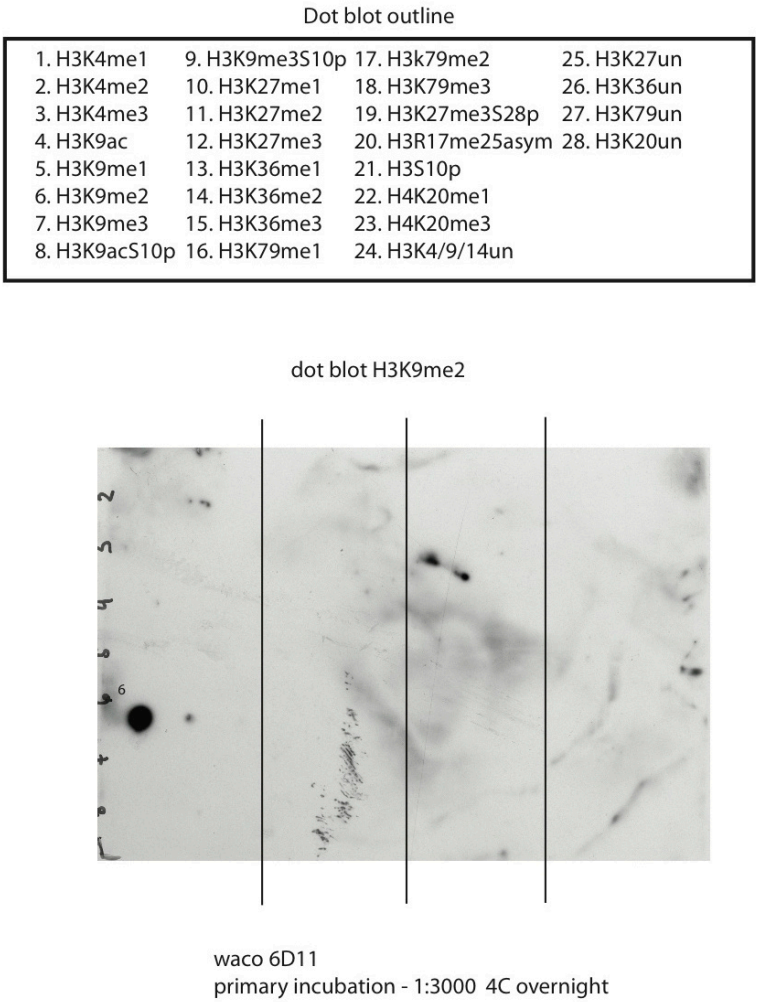


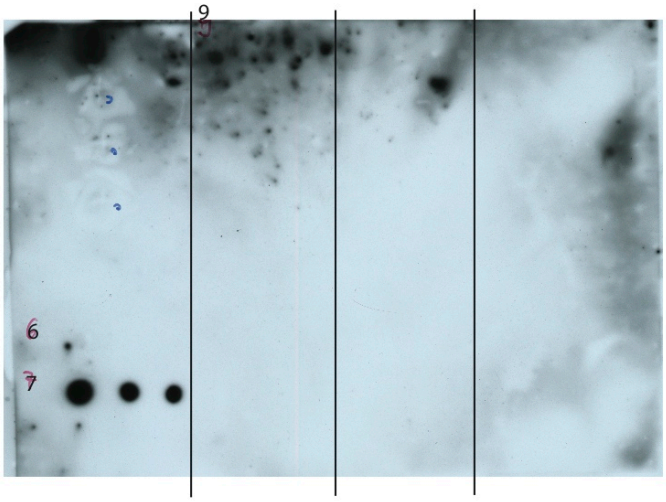
Figure 125 Cross-reactivity western blot for H3K9me2 antibody (WA 309-34839). Westend blot was done in Susan Strome laboratory during collaboration on modENCODE project.

Dot blot outline

1. H3K4me1	9. H3K9me3S10p	17. H3k79me2	25. H3K27un
2. H3K4me2	10. H3K27me1	18. H3K79me3	26. H3K36un
3. H3K4me3	11. H3K27me2	19. H3K27me3S28p	27. H3K79un
4. H3K9ac	12. H3K27me3	20. H3R17me25asym	28. H3K20un
5. H3K9me1	13. H3K36me1	21. H3S10p	
6. H3K9me2	14. H3K36me2	22. H4K20me1	
7. H3K9me3	15. H3K36me3	23. H4K20me3	
8. H3K9acS10p	16. H3K79me1	24. H3K4/9/14un	

Each peptide was spotted in three amounts, 100, 25, and 10 pmol

dot blot H3K9me3



ab8898 lot: 339901
primary incubation - 1:3000 4C over night

Figure 126 Cross-reactivity western blot for H3K9me2 antibody (WA 309-34839). Westend blot was done in Susan Strome laboratory during collaboration on modENCODE project.